Foundation, Implementation and Evaluation of the MorphoSaurus System

Subword Indexing, Lexical Learning and Word Sense Disambiguation for Medical Cross-Language Information Retrieval

Dissertation
zur Erlangung des akademischen Grades
Doctor philosophiae (Dr. phil.)

vorgelegt dem Rat der Philosophischen Fakultät
der Friedrich-Schiller-Universität Jena

von

Kornél Géza Markó, M.A.

geboren am 26.08.1971 in Stuttgart

Gutachter:

1. Prof. Dr. Udo Hahn (Friedrich-Schiller-Universität Jena)
2. Prof. Dr. Rüdiger Klar (Albert-Ludwigs-Universität Freiburg)
3. Prof. Dr. Rainer Hammwöhner (Universität Regensburg)

Tag des Kolloquiums: 15. Oktober 2008

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Have over 35 years of Health Informatics made Europe healthier ?
(Bryden, 2003)

The scientific discipline of *Medical Informatics* or *Health Informatics* aims at establishing the methodological canon of computer science into the context of health and medicine related data, information and knowledge. Medical Informatics applications are strongly user-centered since health professionals are increasingly facing the problem to deal with large amounts of sensitive data in a time-critical setting. Thus, the imperative of health information is that the *proper knowledge must be delivered to the right person, at the right time, in the right place.*

Obviously, the contribution of Medical Informatics to the health of society cannot be measured easily. Anyway, Bryden (2003) believes that Health Informatics *really* made Europe healthier. But since this statement cannot be proven, he proposes another perspective for the definition of Medical Informatics. It is *"using informatics with the goal of improving the health of society"*.

Practical applications of Health Informatics support a broad range of information processing activities in the health sector, targeting different user groups: *Health professionals* (physicians, nurses, and others) in hospitals, outpatient departments and private practices are mainly interested in recording and communicating patient information, ranging from free-text reports over numeric (lab) data to digital biosignals and bioimages. *Health administrators* who are active in the same institutions,

but also in insurance companies and public bodies, are mostly focusing on structuring data for billing and accounting purposes on the one hand, and for health statistics, epidemiology and prevention on the other. *Biomedical researchers*, both in the field of basic and clinical research, are aiming at an adequate representation and documentation of new biomedical knowledge acquired. Finally, medical faculties, teachers and educators are interested to bring the curricular contents to their students using media-supported didactic techniques such as computer-aided instruction and simulation, also including medical information and education resources for laypersons.

Across these application domains, a major bias is given by the following phenomenon: Creators and consumers of primary data (discharge summaries, pathology reports, data from medical imaging, laboratory values, etc.) are mainly interested in unstructured information: radiologists exchange images and the communication between physicians is mainly based on free text, as well as research papers and didactic textbooks. The production of well organized data repositories at the point of care costs more than it brings in, and so tends to be carried out without the carefulness required. In contradistinction, administrators and epidemiologists need highly dependable aggregated structured data in which details are purposely neglected in order to comply with disease, procedure, drug, or patient classification systems. This need of structured information emerged in the discipline of *medical documentation*, for which controlled vocabularies, medical terminologies or even sophisticated ontologies serve as the connecting link for the accurate exchange of medical data between heterogeneous information systems. Such structured information is then used for morbidity and mortality statistics, and for the delineation of homogeneous patient groups in terms of per capita expenditures which plays a major role for quality management routines. Several subdisciplines of Medical Informatics are directly involved in this challenge:

- *Medical Information Systems* provide the physical and logical data infrastructure for the support of medical documentation.
- *Information Retrieval in Medicine* adapts content indexing and retrieval techniques to the medical domain, and is tightly related to

- *Medical Language Processing* which, finally, studies all facets of the sublanguage used in the communication between health care professionals, as well as the text produced by medical authors.

1.1 Medical Information Systems

Information systems that are deployed in the field of medicine are a collection of computer programs for the organization of medical, administrative and scientific information. They are used for the maintenance of, e.g. patient master data, the archiving of patient-related data and their classification (for example the indexing of diagnoses according to the International Classification of Diseases ICD-10 (2005)), the planning of medical service delivery (clinical pathways) and its billing according to Diagnosis Related Groups (DRGs). Many kinds of heterogeneous information are covered in subsystems, e.g. radiology information systems (RIS), patient data management systems (PDMS), picture archiving and communication systems (PACS), and many more.

Besides such patient-related information systems, which are used by health professionals in their every day life, scientific and other information, which are not directly connected to a patient's infirmity, are used by researchers, health-care managers, and others. For example, the Cochrane Library (Chalmers, 1993) and its national spin-offs provide information of up-to-date, high-quality surveys about the effectiveness of therapeutic interventions for easing the decision process for both practioners and patients .

Health care consumers, on the other hand, often rely on information which can be found in the Web. Many health oriented Internet portals exist and are accessible by using search engines. Many people, especially patients affected by certain chronic diseases, exchange information in discussion forums and organize themselves in virtual communities.

Regardless of aiming at medical experts or laypersons – a huge variety of health related resources are available electronically, either as provided within hospital information systems and other intranets, or publicly accessible via the Web. As for

other domains, the amount of medical information grows exponentially, and hence, in order to manage information explosion, great importance is attached to the development of effective tools for the retrieval of specific data.

1.2 Information Retrieval in Medicine

Information retrieval (IR) is a broad interdisciplinary field covering information and computer science, linguistics, semiotics, and librarianship (Baeza-Yates & Ribeiro-Neto, 1999). It deals with the art and science of searching for information in documents, or searching for documents themselves. Two fundamental characteristics distinguish information or document retrieval from the search within more or less simple databases. Firstly, the information need of searchers is vague and can not be formally expressed. Secondly, the information retrieval system stores unstructured data such as natural language texts, and hence, does not 'know' anything about the content. Search engines are designed to find those relevant documents of a collection, which 'somehow' best fit to a particular user query – as selective as possible.

Information and document retrieval plays a crucial role in medical document archiving systems, simply because of the huge amount of data generated in that domain, whether in a health care or in scientific research. Clinical document collections are usually very large and dynamic, with estimates ranging, for a single site, on the order of millions of documents in total, and hundreds to thousands new documents being added every day. The same dynamics can be observed for biomedical publications increasingly electronically available on the Web. PubMed,[1] the interface to the MEDLINE database which is a service of the U.S. National Library of Medicine, gives access to over 16 million citations of life science journals for biomedical articles. MEDLINE is growing at a double-exponential pace. More than three million publications were published in the last five years alone. Moreover, the number of publications indexed in MEDLINE in 2005 was 666,029, i.e. more than 1,800 per day (Hunter & Cohen, 2006). As can be seen in Figure 1.1,[2] since

[1] `http://www.ncbi.nlm.nih.gov/entrez/`, all links last visited in January 2007

[2] `http://www.ncbi.nlm.nih.gov/About/tools/restable_stat_pubmed.html`

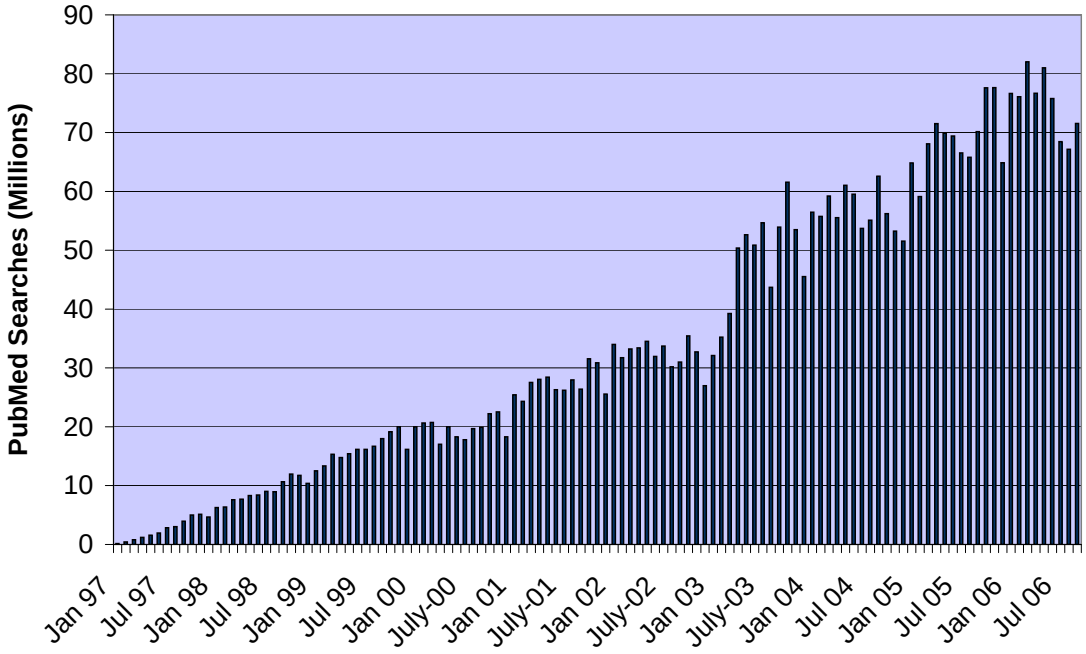

Figure 1.1: Number of PubMed Searches per Month

its release in 1997, the number of requests to the service constantly grew, with now over 70 million searches per month.

The main challenges for the architecture and implementation of document retrieval systems and their underlying search engines are, besides technical issues, inherently linguistic, and additional complexity emerge from the multilingual dimension of information retrieval applied to the medical domain (Hersh, 2002). While clinical documents are typically written in the physicians' native language, searches in scientific databases require sophisticated knowledge of (expert-level) English medical terminology which most non-English speaking physicians do not have. Hence, some sort of bridging between synonymous or, at least, closely related terms from different languages has to be realized to make use of the information these databases or the Web hold.

Furthermore, the user population of medical document retrieval systems and their search strategies are really diverse. Not only physicians, but also nurses, medical insurance companies and patients are increasingly getting access to these resources, with the Web adding an even more scattered crowd of searchers. Hence,

mappings between different jargons and sublanguages are inevitable to serve the needs of such a heterogeneous searcher community. The simplicity of the content representation of the documents, as well as automatically performed intra- and interlingual lexical mappings or transformations of equivalent expressions become crucial issues for an adequate methodology of medical information retrieval.

1.3 An Interdisciplinary Approach

This work is characterized by the challenges of interdisciplinarity of Medical Informatics on the one hand, and Computational Linguistics, on the other. *Medical Linguistics* applies formalisms and methods of general linguistics to the domain-specific medical terminology. In *Medical Language Processing* (MLP), findings of research on *Computational Linguistics* (Allen, 1995; Manning & Schütze, 1999) are adopted for the automatic processing of (spoken or written) medical language. Meanwhile, MLP has been established as a separate and accepted field of research (cf. Spyns (1996) for an overview).

Since the early 70s, remarkable effort has been made in the automatic analysis of medical texts within the *Linguistic String Project* (Sager et al., 1987). However, it is conspicuous that most work has been done in the context of particular areas of application, often along with commercial interests (Lyman et al., 1991). Accordingly, Medical Informatics researchers are not focused on the creation of linguistic theories, rather than this, methods for comprehensive evaluations of MLP systems are proposed, regarding their performance and usability in real world scenarios (cf. Friedman & Hripcsak (1998), Zweigenbaum et al. (1997)). However, although due to other reasons, the demand to evaluate systems for the automatic analysis of language processing systems emerged in a trend that increasingly dominates the domain of Computational Linguistics.

Only recently, remarkable knowledge is transferred between each domains and both communities accrete to one discipline. This development has benefited from the appearance of numerous research groups focusing on text processing for the domains of Biology, Genetics, and Proteomics and there is more and more scientific

work published and conferences held jointly. But still, at least for clinical applications, there is a lack of linguistic knowledge and methodology. In this spirit, this interdisciplinary contribution can further mediate between the two disciplines.

1.4 Overview on this Work

This work proposes an approach which is intended to meet the particular challenges of Medical Language Processing, in particular medical information retrieval. At its core lies a new type of dictionary, in which the entries are equivalence classes of subwords, i.e., semantically minimal units. These equivalence classes capture intralingual as well as interlingual synonymy. As equivalence classes abstract away from subtle particularities within and between languages and reference to them is realized via a language-independent conceptual system, they form an *interlingua*. In this work, the theoretical foundations of this approach are elaborated on. Furthermore, design considerations of applications based on the subword methodology are drawn up and showcase implementations are evaluated in detail.

Starting with the introduction of *Medical Linguistics* as a field of active research in Chapter two, its consideration as a domain separated form general linguistics is motivated. In particular, morphological phenomena inherent to medical language are figured in more detail, which leads to an alternative view on medical terms and the introduction of the notion of subwords. Chapter three describes the formal foundation of subwords and the underlying linguistic declarative as well as procedural knowledge. An implementation of the subword model for the medical domain, the MorphoSaurus system, is presented in Chapter four. Emphasis will be given on the multilingual aspect of the proposed approach, including English, German, and Portuguese. The automatic acquisition of (medical) subwords for other languages (Spanish, French, and Swedish), and their integration in already available resources is described in the fifth Chapter.

The proper handling of acronyms plays a crucial role in medical texts, e.g. in patient records, as well as in scientific literature. Chapter six presents an approach, in which acronyms are automatically acquired from (bio-) medical literature. Fur-

thermore, acronyms and their definitions in different languages are linked to each other using the MorphoSaurus text processing system.

Automatic word sense disambiguation is still one of the most challenging tasks in Natural Language Processing. In Chapter seven, cross-lingual considerations lead to a new methodology for automatic disambiguation applied to subwords.

Beginning with Chapter eight, a series of applications based on MorphoSaurus are introduced. Firstly, the implementation of the subword approach within a cross-language information retrieval setting for the medical domain is described and evaluated on standard test document collections. In Chapter nine, this methodology is extended to multilingual information retrieval in the Web, for which user queries are translated into target languages based on the segmentation into subwords and their interlingual mappings.

The cross-lingual, automatic assignment of document descriptors to documents is the topic of Chapter ten. A large-scale evaluation of a heuristic, as well as a statistical algorithm is carried out using a prominent medical thesaurus as a controlled vocabulary.

In Chapter eleven, it will be shown how MorphoSaurus can be used to map monolingual, lexical resources across different languages. As a result, a large multilingual medical lexicon with high coverage and complete lexical information is built and evaluated against a comparable, already available and commonly used lexical repository for the medical domain.

Chapter twelve sketches a few applications based on MorphoSaurus. The generality and applicability of the subword approach to other domains is outlined, and proof-of-concepts in real-world scenarios are presented.

Finally, Chapter thirteen recapitulates the most important aspects of MorphoSaurus and the potential benefit of its employment in medical information systems is carefully assessed, both for medical experts in their everyday life, but also with regard to health care consumers and their existential information needs.

Chapter 2

A Morphological Perspective on Medical Language Processing

Practical tasks of Medical Linguistics are the development and implementation of algorithms, which render services customized for the medical sublanguage. Typical usages are, e.g. spelling correction software and other programs to aid physicians with the generation of documents. Computer-aided classification of diagnoses and automatic text categorization assign terms from a controlled vocabulary to medical documents (Aronson et al., 2000; Sebastiani, 2002). Information Retrieval (IR) systems usually give access to huge document collections, either stored in clinical information systems or publicly available (Hersh & Donohoe, 1998; Eichmann et al., 1998; Volk et al., 2002). Information Extraction (IE) concerns the automatic processing of unstructured, textual data aiming at acquiring factual, structured knowledge from these documents (cf. Hahn et al. (2002b) for the analysis of pathology findings, and Friedman et al. (1994) for radiology reports). IE systems proved to be useful, e.g. for the automatic identification of clinical findings suspicious for tuberculosis (Jain et al., 1996) or breast cancer (Jain & Friedman, 1997). Finally, Text Mining systems, prevailing in the biomedical domain, are implemented for the generation of new knowledge, which implicitly exists across different documents in (usually) huge document collections (Feldman et al., 1999; Liu & Friedman, 2003; Shatkay & Feldman, 2003; Nenadić et al., 2003; Daraselia et al., 2004).

It has been observed that medical language shows less syntactic variation and complexity than general language as found, e.g. in newspapers, narratives, etc. (Campbell & Johnson, 2001; Friedman & Hripcsak, 1999). However, it is still controversial whether off-the-shelf NLP tools can be effectively ported or adopted to the needs of MLP. At least for the syntactic analysis of medical texts, evidence has been found that statistical NLP methods can be used in a straightforward manner (Hahn & Wermter, 2004; Wermter & Hahn, 2004). However, the most obvious contrast of domain-specific sublanguages to general language is the use of a profoundly different vocabulary together with a highly complex morphology.

2.1 Medical Linguistics

Typically, the word pool of a language is estimated to range between 200,000 and 500,000 words, depending whether domain-specific terminology is included, or not. The Oxford English Dictionary (Simpson & Weiner, 1989) is generally regarded as being the most comprehensive dictionary of the English language and includes more than 500,000 main entries, both for present and past English. General language dictionaries contain between 100,000 and 150,000 entries. In comparison, medical dictionaries additionally include at least 50,000 words (Taber, 2005; Roche, 2003).

Natural language is furthermore characterized by morphological processes, which tend to alter the literal appearance of the lexical items but let the meaning core of these entities largely unchanged. Such morphological variants can generally be described as concatenations of basic lexical forms (stems) with additional substrings (affixes). The diversity of morphological processes varies between languages, with English known as a morphologically 'poor' language, while most others are much more diverse. Evidence for this statement comes from a large variety of highly inflectional and/or agglutinating languages such as Finnish (Jäppinen & Niemistö, 1988), Hebrew (Choueka, 1990), Slovene (Popovič & Willett, 1992), Turkish (Ekmekçioglu et al., 1995), Swedish (Hedlund et al., 2001), German (Schulz et al., 2002) or Hungarian (Tordai & de Rijke, 2005), not to mention major Asian languages such as Japanese, Korean and major dialects of Chinese.

The medical sublanguage reveals even more complexity. Ancient Greek doctors usually used metaphors to describe body parts and diseases. For example, the part between the stomach and the small intestine has a length of twelve fingers, and therefore was called *"dodekadaktylos"*, later adopted to Latin as *"duodenum digitorum"*. Some Greek terms were not simply translated, instead of this, many of them were replaced, for example the Greek word *"spondylos"* (the bone element of the spine) was replaced by *"vertebra"* (literally, *"the rotating"*). Latin generally became the prevailing language for science, but terms for diseases, in contradistinction to anatomy, were still composed of Greek roots. That is the reason why anatomists use the word *"vertrebra"*, whilst clinicians use *"spondylitis"* when they refer to an inflammatory disorder of a vertebra. Accordingly, the Latin word for kidney is *"ren"*, the inflammation of the kidney is called *"nephritis"* (derived from the Greek stem *"nephr"*).

These days, medical terminology is characterized by a typical mix of Latin and Greek roots with the corresponding host language, often referred to as *neo-classical compounding* (McCray et al., 1988), e.g. in words such as *"neuroencephalomyelopathy", "glucocorticoid", "pseudohypoparathyroidism"*. Morphologically rich languages (e.g., German) tend to conflate these terms, moreover, with host language terms, resulting in longer single-word compounds such as *"Gastrointestinaltrakt", "Kortikoidmedikation"*, etc. This also results in a high amount of synonymous terms, which express the differences of experts and laypersons terminology.

2.2 Morphological Processes

In linguistics, morphology is the field of research that studies the formal properties and internal structure of words. Words are composed of morphemes, which are commonly defined as the minimal units of meaning. Three kinds of morphological processes are generally distinguished, viz. inflection, derivation and composition inherent to so-called agglutinative (sub-) languages.

- *Inflection* adds number, gender or case information to nouns (e.g., *"patient⊕s"*)[1], or number, person, and tense information to verbs (e.g., *''injur⊕es"*, *"injur⊕ed"*, *"injur⊕ing"*). These modifications are typically motivated by syntactic considerations, thus, the lexical sense of the word stem is combined with the grammatical function of the affixes.

- *Derivation* covers different phenomena. A derivational affix may simply affect the part of speech without any semantic implication (e.g. *"patient with a severe⊕_ injur⊕y"* is a *"severe⊕ly injur⊕ed patient"*). Often, minor changes of the semantic interpretation of the derived form relative to the basic one occur (e.g., *"search⊕er"* denotes 'someone who *"search⊕es"*).

- *Composition*, finally, combines several basic lexico-semantic units to form a composite one. In contrast to English where nominal compounds surface as complex noun phrases (Levi, 1978), e.g., *"femoral neck fracture"*, agglutinative languages such as German (Toman, 1987) build up complex single-word compounds (e.g., in the translation *"Ober⊕schenkel⊕hals⊕bruch"*).

Morphological analysis is concerned with the reverse processing, i.e., deflection (or *lemmatization*), dederivation and decomposition. The goal is to map all occurring morphological variants to some canonical base form(s), e.g., *"injur"* in one of the examples from above.

2.3 Morphology in Medical Terminology

In order to collect empirical evidence for the question whether morphological analysis of complex word forms is really an urgent need, Schulz & Hahn (2000) conducted the following experiment: In a random selection of 100 pathology reports (average token count 147.9 per report) they found 895 occurrences of different domain-specific compounds. They then matched these 895 forms against all words contained in a machine-readable version of a comprehensive German-language medical dictionary,

[1] '⊕' denotes the string concatenation operator.

Language	Compound	Segmentation
English	Pseudohypoparathyroidism	Pseudo⊕hypo⊕para⊕thyroid⊕ism
	Proctosigmoidoscopy	Proct⊕o⊕sigm⊕oid⊕o⊕scop⊕y
	Arterionephrosclerosis	Arteri⊕o⊕nephr⊕o⊕scler⊕osis
German	Kryostatschnittverfahren	Kryo⊕stat⊕schnitt⊕verfahr⊕en
	Fibroblastenproliferation	Fibro⊕blast⊕en⊕prolifer⊕ation
	Koronarangioplastie	Koron⊕ar⊕angio⊕plast⊕ie
Portuguese	Electrocardiográfia	Electr⊕o⊕cardio⊕gráfi⊕a
	Imunodeficiência	Imun⊕o⊕defici⊕ência
	Esofagocardiomiotomia	Esofag⊕o⊕cardio⊕mio⊕tomia
Spanish	Hipersomatotrófico	Hiper⊕somat⊕o⊕tróf⊕ico
	Postpericardiotomía	Post⊕peri⊕cardio⊕tom⊕ía
	Hepatoesplenomegálica	Hepat⊕o⊕esplen⊕o⊕megal⊕ica
French	Cholécystographie	Cholé⊕cyst⊕o⊕graph⊕ie
	Polychimiothérapie	Poly⊕chimi⊕o⊕thérap⊕ie
	Épidermodysplasie	Épi⊕derm⊕o⊕dysplas⊕ie
Swedish	Blindtarmssjukdomar	Blind⊕tarm⊕s⊕sjukdom⊕ar
	Övergångsepitelcancer	Över⊕gång⊕s⊕epitel⊕cancer
	Hjärnnervstumörer	Hjärn⊕nerv⊕s⊕tumör⊕er

Table 2.1: Medical Nominal Compounds in Different Languages

the *"Pschyrembel"*.[2] The retrieval process was based on exact string match. As a result, 400 out of these 895 compounds were not found in the dictionary. This reflects the enormous productivity of medical language leading to a large number of *ad hoc* compounds. A number of examples in different languages are given in Table 2.1.

Analyzing the rubrics of the English-language coding system ICD-9-CM (cf. ICD-10 (2005) for its successor), Schulz & Hahn (2000) found a considerable num-

[2] *Pschyrembel Klinisches Wörterbuch*, Walter de Gruyter. Its whole text corpus contains more than 100,000 different entries.

ber of nominal compounds (cf. the English terms in Table 2.1), thus indicating that this phenomenon is by no means restricted to the German language only. Generalizing from this study, the hypothesis is confirmed that accounting for complex morphological phenomena is highly rewarded in medical language processing.

For the medical terminology, morphological complexity further increases, in structural terms and independent of particular languages (Ingenerf, 1997; Rector, 1999). For example, by means of composition, the basic word forms *"leukocyte"* and *"[H]em⊕o"* join into *"Leuk⊕em⊕ia"*, with a tricky omission of the starting character of *"Hemo"*, and the use of *"ia"* as suffix. Other unsystematic modifications can often be observed in clinical findings, where ad-hoc compounds appear frequently. They are invented on the spot and may never be used again. In many cases, the meaning of compounds can not be derived by their constituents (as, e.g. in *"antibodies"*). Noun compounds or multi-word terms co-exist with Latin noun phrases and the use of Latin and Greek roots results in a high amount of synonymous terms, which also reveal the differences of experts' and laypersons' terminology. In addition, different orthographic variants of the same word can be observed for Latin and Greek loanwords (e.g. *"collum uteri"* vs. German *"Uteruskollum"* or *"leucocyte"* and *"leukocyte"* in English.

Acronyms also play a crucial role in medical documents, both in clinical reports, as well as in scientific publications. Actually, the extensive use of acronyms and abbreviations in the biomedical community has been highly criticized (Rowe, 2003). It is estimated that the number of unique acronyms in scientific publications related to biomedicine is increasing at a rate of approximately 11,000 per year, whilst the number of definitions associated with them is growing at approximately four times that rate (Wren & Garner, 2002). Since 36% of all acronyms in MEDLINE are associated with more than one definition and, conversely, up to 10% of definitions are associated with more than one acronym, disambiguation techniques are inevitable necessary in order to account for them properly.

2.4 Morphology in Information Retrieval

In a common free-text information retrieval environment, the search for a particular document is based on an (exact) pattern matching operation between the query term(s) and the document terms. Therefore, a query term such as *"leukocytes"* retrieves all those documents in which this query term occurs literally. On the other hand, documents containing the singular form *"leucocyte"*, the adjective *"leukocytic"*, or the compound term *"leukemia"* cannot be found. In order to account for morphological variations of terms, three basic possibilities arise in a free-text retrieval system:

1. Enumerate all morphological variants of a query term and combine them, either manually or automatically. Afterwards, combine the resulting variants in a disjunction, such as in *"Leukocyte" OR "Leukocytes" OR "Leukocytic" OR "Leukemia" OR*... . Then let the system perform exact matches with corresponding document terms.

2. For a given query term, a truncation operator (such as '*', or '%' in relational database systems) is applied to the longest common substring of all possible morphological variants, e.g., *"leuk**". The system will then perform a partial string match of this truncated query term and all document terms whose leftmost substring is identical with *"leuk"*, while the remainder can be any arbitrary string. Such a mechanism mimics linguistically based morphological computations by simple string processing approximation.

3. Determine morphologically motivated base forms of query terms and document terms, e.g., *"leukocyt"*, and let the system automatically cope with morphological variants using a considerable amount of linguistic knowledge. The matching between query and document terms is then performed by the system based on these system-determined variant sets.

The first approach often yields incomplete coverage, especially in subdomains as the medical one, due to missing variants, even for linguistically well-trained human

searchers of the particular domain. Therefore, this alternative leads to an incomplete search for relevant documents (low recall). Even worse, for morphologically rich (sub-) languages which include single word compounding in their word generation process, this approach is not feasible at all. Contrarily, the second solution tends to overgenerate, and therefore finds irrelevant documents for a given query (low precision), producing many unintended matches, since the matching process is entirely underconstrained (e.g. querying *"aid*"* which would also match *"AIDS"*.

Considering the third suggestion, different methodologies for the automatic analysis of morphological variants have to be distinguished in order to assess potential benefits or drawbacks for document retrieval systems.

2.5 Medical Morphological Analysis

For information retrieval, the most common approach to morphological analysis is based on *stemming*, i.e., conflating different morphological variants to a single formal stem. Typically such algorithms (e.g., the Lovins stemmer (Lovins, 1968) or the Porter stemmer (Porter, 1980)) refrain from using dictionary information and are solely based on simple string processing routines. Their principal way of operation consists of removing inflectional endings (e.g., plural or genitival or tense suffixes) or derivational suffixes, including some recoding transformations. Some of them, e.g., the Lovins stemmer, follow a one-pass strategy based on right-to-left longest matching plus recoding. Others, e.g. the Porter stemmer, employ an iterative multi-pass approach. In fact, there has been some controversy about their contribution to improve the effectiveness of document retrieval systems (Harman, 1991; Krovetz, 1993; Hull, 1996; Kantrowitz et al., 2000; Tomlinson, 2001; Braschler & Ripplinger, 2004; Tordai & de Rijke, 2005).

The key issue for quality improvement seems to be rooted mainly in the presence of some kind of dictionary, i.e., a list of content words in some agreed-upon basic lexical format plus, possibly, additional linguistic information concerning parts of speech, gender, number, tense, mood, semantic relations, etc. Empirical evidence has been brought forward that inflectional and/or derivational stemmers augmented

by (machine-readable) dictionaries perform substantially better than those without access to lexical repositories (Krovetz, 1993).

In addition, the above-mentioned stemming algorithms and their many variants benefit from the limited suffix set and rather simple formation rules underlying English inflection. When turning to other languages, e.g., French, Italian, Spanish, or German, no comparable algorithmic standard yet exists. Many of these languages exhibit a much richer inventory of inflectional suffixes, and also their structural combination is more complex. Evidence for this statement comes form a large variety of highly inflectional and/or agglutinating languages.

Morphological complexity further increases, in structural terms and independent of particular languages, when one looks at derivation and composition (for a survey of German, cf. Toman (1987), for English composition, cf. Levi (1978)). There have already been observations on the crucial status of compounds for information retrieval and the problems they cause (Jäppinen & Niemistö, 1988). This becomes particularly pertinent for the medical domain where a large number of established terms with a considerable morphological complexity exist.

It is worth mentioning that pessimism has been expressed with respect to a full semantic interpretation of medical compounds (McCray et al., 1988). However, several approximations have already been proposed. The earliest approach to deal with medical terminology by way of morphological analysis is due to the work of Pratt & Pacak (1969). Their approach transforms semantically equivalent adjectival and nominal forms by employing simple suffix trees and transformation rules for recoding morphologically reduced forms. Such transformations succeed if a recoded form is matched with an entry in the Systemized Nomenclature of Pathology (SNOP, which later evolved into SNOMED, the Systematized Nomenclature of Medicine (Côté et al., 1993; CT, 2004)). Using a defined vocabulary for term normalization in the medical domain is also reported more recently, e.g. in the work of Zeng & Cimino (1996) and Kornai (2004).

Follow-up studies by Pacak and Norton (Pacak et al., 1980; Norton & Pacak, 1983) not only determined a preferred normalized form for several morphological variants but rather computed paraphrase and other semantic relations (such as loca-

tive, causative, etc.). These are implicitly denoted by complex medical compound nouns and can be made explicit by breaking compounds up into their constituent parts. The distributional patterns Pacak and Norton suggest are based on four top-level conceptual categories which are directly derived from SNOP/SNOMED codes (*viz.* topography, (medical) morphology, etiology, and function). A major limitation of this work, however, is the restriction of the decompositional analysis to inflammatory processes (indicated by the suffix *"-itis"*) or to surgical procedures (indicated, e.g., by the suffixes *"-ectomy"* or *"-plasty"*) only. In a similar vein, Dujols et al. (1991) treat *"-osis"* endings only, though in a slightly more sophisticated manner. These restrictions are somewhat weakened in the work of Wolff (1984) both in terms of a larger number of Greco-Latin suffixes being covered, as well as more general compositional patterns of Neo-Latin compounding. However, the conceptual categories she employs refer to the subclass coding principles specifically employed in the LSP context, the Linguistic String Project (for an overview, cf. Sager et al. (1994)), rather than to the conventional SNOP/SNOMED-style nomenclature.

A lot of this work is characterized by a mix of isolated data structures (e.g., suffix trees) and various procedural heuristics (longest match from the right, floating *"o"* insertion as in *"cyst⊕o⊕lith⊕ectomy"* vs. *"cyst⊕ectomy"*, etc.). In an attempt to formulate the principles of medical word segmentation in a formally rigid, almost language-independent framework, Wingert (1977) chose an automaton-based specification for morphological analysis in terms of augmented transition networks. To this end he proposed a set of 255 cascading rules to capture the combinatorial regularities of different morpheme classes and, similar to Pratt & Pacak, refers to the entries of the SNOP nomenclature in order to exploit semantic information from the medical domain (Wingert, 1985).

As an alternative, remarkable progress has already been made by Yarowsky & Wicentowski (2000) and Goldsmith (2001) in the fields of supervised and unsupervised acquisition of morphological units (i.e., stems and affixes), including the alignment of potential stem changes due to inflection. Unfortunately, none of the reported systems are capable of performing noun decomposition, which is essential for the analysis of medical terminology.

Much more sophisticated linguistic and conceptual knowledge is employed in more recent work on medical morphology. Lovis et al. (1997), Baud et al. (1998) and Baud et al. (1999) use finite-state technology for the decomposition of complex terms into semantically non-decomposable segmentation units they refer to as *morphosemantemes*. A lot of the power of their approach derives from the fact that the conceptual correlates of these morphosemantemes no longer refer to flat SNOP/SNOMED-style categories but rather are formulated in GRAIL, a highly expressive deductive terminological knowledge representation language within the GALEN framework (Rector et al., 1997). In order to isolate a morphosemanteme, composite concepts are dissected to their medically plausible conceptual core, using the knowledge encoded in GRAIL.

Baud et al.'s approach fully depends on the terminological coverage of the medical domain by GRAIL which, as any of *deep* knowledge approaches, hardly scales up to reasonably sized, practically-to-use knowledge bases.

It is interesting to observe that none of the above-mentioned proposals make use of the state-of-the-art methodologies for morphological analysis in natural language processing, *viz.* chart-based approaches in the (early) eighties (Kay, 1980), or the model of two-level morphology as originally formulated by Koskenniemi (1984) and lucidly described in Sproat (1992). The reason might be that these pure NLP methodologies still pose too strong requirements on their linguistic resources (e.g., two-level morphology requires elaborated and complete stem and suffix lexicons) and are also too rigid with respect to well-formedness of their input. So far, major efforts have been directed at deflection only, with minor attention being paid to derivational (Russell et al., 1986; Trost, 1993) or compositional morphology (Black et al., 1991; Karttunen et al., 1992). Even worse, some languages such as German pose particularly problems to a two-level approach because of contextual alteration dependencies within words such as umlauts or participles (cf. Trost (1990) and Schiller & Steffens (1991) for an overview), not to mention the problem of mixed-language input, as evidenced by Neo-Latin compounding in medical terminology.

2.6 MorphoSaurus

Since the year 2000 a unique and powerful medical language tool has been developed in the Department of Medical Informatics at the University Hospital in Freiburg, Germany, in cooperation with the Language and Information Engineering Lab at Jena University, Germany, and the Paraná Catholic University in Curitiba, Brasil. The basic component of the system, a medical thesaurus that roughly consists of morphemes, led to the name MorphoSaurus (an acronym for MorpHeme theSaurus). It provides a methodology for morphological analysis that accounts for (a) all three basic morphological processes, i.e., inflection, derivation, and composition, and (b) the combination of Greek, Latin, and a particular host language (in the current implementation English, German, French, Spanish, Portuguese, and Swedish). Unlike approaches which are purely driven by considerations of general natural language processing, the methodology proposed here focuses on medical Cross-Language Information Retrieval (Markó et al., 2004b; Hahn et al., 2004a; 2005b; Markó et al., 2005c; Daumke et al., 2005b), additionally considering other multilingual applications such as terminology mapping (Markó et al., 2003; Markó et al., 2004a; Hahn et al., 2004b; Markó et al., 2006c) or lexicon mapping (Markó et al., 2006a; 2006b). This focus has concrete implications for (c) the choice of the fundamental unit of morphological analysis, as well as (d) the way how these units are semantically related within, but also across languages. Though the name MorphoSaurus is derived from a kind of morpheme-based thesaurus, the notion of a lexical unit is slightly broader than the linguistic definition of a morpheme, but clearly narrower than full forms of words. This led to the introduction of so-called *subwords* (Schulz & Hahn, 2000; Hahn et al., 2001; Markó et al., 2005a; 2005d; Hahn et al., 2005b; Schulz et al., 2006).

Chapter 3

Subword Model

The conventional view on human language is word-centered, at least for written language where words are clearly delimited by spaces. It builds on the hypothesis that words are the basic building blocks of phrases and sentences. In syntactic theories, words constitute the terminal symbols. Therefore, it appears straightforward to break down natural language to the word level. However, looking at the sense of natural language expressions, evidence can be found that semantic atomicity frequently does not coincide with the word level, which bears methodical challenges even for pretendedly 'simple' tasks such as tokenization of natural language input (Grefenstette & Tapanainen, 1994). As an example, considering the English noun phrase *"high blood pressure"*, the word limits reflect quite well the semantic composition, whereas this is not the case in its literal translations *"verhoogde bloeddruk"* (Dutch), *"högt blodtryck"* (Swedish) or *"Bluthochdruck"* (German). Especially in sublanguages such as the medical one, atomic senses are encountered at different levels of lexical granularity. An atomic sense may correspond to word stems (e.g., *"hepat"* referring to *"liver"*), prefixes (e.g., *"anti-"*, *"hyper-"*), suffixes (e.g., *"-logy"*, *"-itis"*), larger word fragments (*"hypophys"*), words (*"spleen"*, *"liver"*) or even multi-word terms (*"yellow fever"*). The possible combinations of these word-forming elements are immense and ad-hoc term formation is common. As a consequence, a high coverage of a domain-specific lexicon can only be expected if lexical units are restricted to units of atomic senses, which then can be used as building blocks for composed

terms at any level of granularity.

Identifying atomic sense units from texts in order to achieve a basis for the (lean) semantic interpretation of natural language texts is an important requirement for many applications in the fields of document retrieval, information extraction, and text mining.

3.1 Semantic Atomicity

In linguistic theories, a sequence of characters are regarded as semantically atomic if the sense conveyed (in a given language and a given domain context) is not univocally derivable from the senses of its constituents.[1] The constituents of words are morphemes, and they are tied together by word-forming operations such as inflexion, derivation and composition. For instance, *"neurosis"* is the result of linking *"neur"* (nerve) with *"osis"* (disease). However, *"neurosis"* is not really a disease of nerves (at least in modern scientific medicine). As a consequence, the derivation *"neurosis"* would be considered an atomic lexical unit.

Lexical units may have multiple senses (homonymy, in a broad sense) and one sense can be expressed by different surface forms (synonymy). Although domain specific terminologies are constructed in order to control the use of a specialized language and to avoid ambiguous expressions, non-standardized terminology is widely used in any domain. For instance, *"molar"* has a completely different sense in obstetrics (*"molar pregnancy"*) as in lab medicine (*"molar mass"*), or in dentistry (*"fractured molar"*). The meaning of the stem *"head"* in *"headache"* is different from the ones in *"head of femur"* or *"head of department"*. *"Operation"* means *"surgical procedure"* in the medical domain, as opposed to different senses in mathematics or business. In such cases, the local context of the word in focus generally helps selecting the right sense. Furthermore, the restriction to a well-defined domain (e.g.

[1] Many semantic theories are still controversially discussed by different scientific disciplines such as philosophy, cognitive science, linguistics, and information science. In this work, the *sense* of a linguistic expression is defined by the *mental construction* that is associated with this utterance, rather than to concrete objects in the world (Eco et al., 1988).

clinical medicine) allows us to ignore word senses which are definitely outside that domain (e.g. *"head"* as the role of a word in grammar theory).

Besides ambiguity, lexical units may have overlapping senses. Quasi-synonymy relations may hold between terms of different languages (Latin *"caput"* vs. English *"head"*) or different levels of erudition (*"belly"* vs. *"abdomen"*). Complete identity in sense (strict synonymy) which holds throughout all possible uses of a word is rare.

In order to establish classes of synonymous expressions, clear commitments to the context in which the expressions can be regarded as synonyms have to be made, *viz.* defining the *domain context*. Secondly, an agreement has to be found on a sense deviation tolerance which is still compatible with the formal properties of an equivalence relation, *viz.* reflexivity, transitivity, symmetry: If one considers *"disease"* as a synonym of *"illness"* and *"illness"* as a synonym of *"sickness"*, then *"disease"* and *"sickness"* are synonyms, as well. The tolerance depends also on the relevance of subtle sense distinctions in the chosen domain context. In the domain of clinical medicine, e.g., *"neoplasm"*, *"cancer"* and *"carcinoma"* would hardly be considered synonyms but a different decision may, however, be taken in another domain. A counterexample would be to equalize *"excis-"*, *"remov-"* and *"-ectom-"* in a domain of general medicine, neglecting subtle distinctions of surgical techniques.

Translation is a special case of synonymy in which words of different languages are sense-linked to each other. In this case, equivalence can be defined as well, e.g. consisting of English *"disease"* and *"illness"*, German *"Krankheit"*, Spanish *"enfermedad"*, French *"maladie"*, Swedish *"sjukdom*, as well as Portuguese *"doença"*.

Not only the grouping of lexical units into synonymy classes, but also their proper delimitation depends on the domain context. *"Leukemia"*, e.g., literally means *"white blood"*, and *"neurosis"* literally means *"nerve disease"*. This may be plausible in a historic view on medicine, but it provides an inaccurate description when related to modern medicine. Thus, a composite sense may be ascribed in the historic context, and an atomic one in the present one.

Within the MorphoSaurus framework, in order to represent atomic senses of lexical units, a semantic layer is defined, which is made of language-independent unique identifiers, so called MorphoSaurus identifiers (shortly, MIDs). These

symbols refer to all lexical items that cover the same meaning, in all languages considered. Equivalence classes can roughly be compared to concepts in thesauri, such as synsets in WordNet (Fellbaum, 1998) or, in the medical domain, concept unique identifier (CUIs) in the Metathesaurus of the Unified Medical Language System UMLS (2005), an umbrella system which currently combines more than one hundred heterogeneous medical terminology systems. The most important ones are available in several languages, e.g. the *International Classification of Diseases (ICD)* (ICD-10, 2005), the *Medical Subject Headings* MeSH (2005), etc.

However, there are two major differences between MIDs and WordNet synsets or UMLS CUIs: Firstly, MIDs can represent disjunctions of different senses. This is the case when ambiguous lexical units are addressed. To restate the example from above, the disjunction of the different senses of *"molar"* is represented by one MID, and each of the non-ambiguous senses by another MID. Secondly, all lexical units which are assigned to one MID must be fully interchangeable. For example, {*head, caput, cabec, cabez, cefal, cephal* } would not be a proper reference for one MID, since *"head"* (in the example denoting a relative anatomical location) has additional senses, at least in a domain context which includes the meaning of *"head"* as a person.

A different view on MIDs is to regard them as non-ambiguous words of an interlingua, since each synonym class is uniquely identified by one MID. This perspective emphasizes the preference of representing lexical meaning abstracting away from the variety of human language, an exercise that must not be mistaken for the construction of concepts or classes in a domain ontology (cf. Hirst (2004) for the relationship between lexicons and thesauri to ontologies).

3.2 Morpho-semantic Indexing

A subword is the minimal semantic constituent of a domain-specific term. Its defining property is that its sense is *not* composite. This rules out, for instance, to consider *"hepatitis"* a valid subword because its sense can be derived from its constituents, in contradistinction to, e.g., *"hypophysis"* (composing the senses of its

components *"hypo"* and *"physis"* does not lead to the proper sense of *"hypophysis"*, i.e. *"hypo⊕phys⊕is"* would be semantically underdeterminate). Subwords can appear as word stems, (proper) prefixes and suffixes, infixes, or invariants.

- Subword stems (ST), like *"gastr"*, *"hepat"*, *"enferm"*, *"diaphys"*, *"head"* are the primary content carriers in a word. They can be prefixed, linked by infixes, and suffixed, some of them may also occur without affixes.
- Prefixes (PF), like *"de-"*, *"re-"*, *"in-"*, *"anti-"* , *"hyper-"* precede a stem or another prefix.[2]
- Proper Prefixes (PP) such as *"peri-"*, *"hemi-"*, *"down-"* are prefixes that themselves cannot be prefixed.
- Infixes (IF), like *"-o-"*, in *"gastr-o-intestinal"*, or *"-r-"*, in *"hernio-r-rafia"* are used as a (phonologically motivated) glue between stems.
- Suffixes (SF) such as *"-a"*, *"-io"*, *"-ion"*, *"-tomy"*, *"-itis"* follow a stem or another suffix.
- Proper Suffixes (PS) (e.g. verb endings such as *"-ing"*, *"-ieron"*, *"-ão"*, *"-iésemos"*) are suffixes that cannot be suffixed.

The classification of subwords like *"-logia"* or *"-itis"* as suffixes may be controversial. As a rule of thumb, the criterion for stems is that they do not require any other stem in order to build well-formed words.

All these lexeme types are used for segmentation of inflected, derived and composed words, taking into account their compositional constraints. In contrast,

- Invariants (IV), like *"ion"*, *"gene"*, many proper names as *"aspirin"* and acronyms such as *"WHO"* or *"AIDS"*

coincide with words and are not allowed as word parts. In most cases, these are short words which would cause artificial ambiguities if they were made available as possible constituents in the deconstruction of complex words.

[2]E.g. in *"hemi⊕an⊕opsia"* the prefix *an* is prefixed by *"hemi"*.

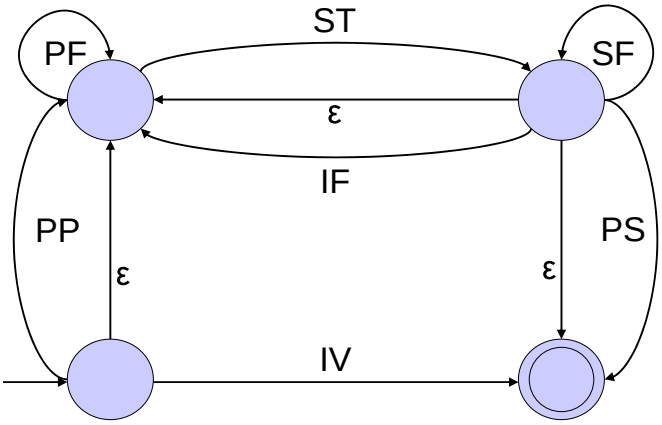

Figure 3.1: Subword Model

Figure 3.1 depicts the subword model in terms of a finite-state automaton. Consequently, a word optionally starts with a (proper) prefix, followed by at least one stem (which can be combined with others, separated by optional infixes or additional pre- and suffixes) and ends with (proper) suffixes, which are optional as well.

3.2.1 Subword Lexicon

Let $\mathcal{S} := \{gastr, hepat, enferm, de, anti, itis, \ldots\}$ be the set of lexical items at the subword level and $\mathcal{T} := \{PP, PF, ST, IV, SF, PS\}$ denote the subword types, as described above. Furthermore, let $\mathcal{M}$ contain the set of equivalence class symbols (MIDs). By convention, elements of this set are annotated with #, followed by the literal entry of an unambiguous English subword or the original, ambiguous subword: $\mathcal{M} := \{\#gastr, \#liver, \#inflamm, \ldots\}$. With $\mathcal{L} := \{EN, GE, FR, SP, PT, SW\}$ referring the set of languages under consideration[3] (English, German, French, Spanish, Portuguese and Swedish, respectively) and $\mathcal{D} := \{ClinicalMedicine, Biology, Chemistry, \ldots\}$ the domain context, a lexical entry is defined by being a member of the lexicon $\mathcal{LEX}$, i.e. the set:

[3]The language attribute refers to the real-world occurrence of lexemes, including common foreign words. This means, e.g., that English lexemes which commonly occur as foreign lexemes in a certain domain (e.g. *"feedback"*) or frequent acronyms which are derived from English long forms (*"WHO"*) are considered lexemes of the respective host language.

$$\mathcal{LEX} \subset \mathcal{S} \times \mathcal{T} \times \mathcal{M} \times \mathcal{L} \times \mathcal{D}$$

If no meaning is assigned to a subword, it is a *stop entry*, having only a grammatical function, such as auxiliary verbs or inflection endings. In this case, the MID attribute is be empty (ε).

The following are some typical examples of subword lexicon entries, their lexical attributes and implicit lexical relations (with $l_{1,\ldots,n} \in \mathcal{LEX}$, and $d_{1,2} \in \mathcal{D}$):

- Synonymy: The suffixes *"-itic"* and *"-itis"* have the same meaning as *"inflammation"*.
  $l_1 = (inflamm, ST, \#inflamm, EN, d_1)$
  $l_2 = (itic, SF, \#inflamm, EN, d_1)$
  $l_3 = (itis, SF, \#inflamm, EN, d_1)$

- Translation: The German stem *"entzünd"* (transliterated to *"entzuend"*) and the French suffix *"-ite"* denote the same sense as the English stem *"inflamm"*.
  $l_1 = (inflamm, ST, \#inflamm, EN, d_1)$
  $l_4 = (entzuend, ST, \#inflamm, GE, d_1)$
  $l_5 = (ite, SF, \#inflamm, FR, d_1)$

- Stop entries: The word *"era"* is an English noun, but an auxiliary verb in Spanish and Portuguese.
  $l_6 = (era, ST, \#era, EN, d_1)$
  $l_7 = (era, IV, \varepsilon, SP, d_1)$
  $l_8 = (era, IV, \varepsilon, PT, d_1)$

- Quasi-synonyms: The word *"sildenafil"* and the name *"viagra"* can be considered synonyms in clinical medicine (d_1), but not in pharmaceutical industry (d_2).
  $l_9 = (sildenafil, ST, \#sildenafil, EN, d_1)$
  $l_{10} = (viagra, IV, \#sildenafil, EN, d_1)$
  $l_{11} = (sildenafil, ST, \#sildenafil, EN, d_2)$
  $l_{12} = (viagra, IV, \#viagra, EN, d_2)$

Table 3.1 shows a minimal lexicon for English (top-left) and German (top-right).

English Subword Lexicon	**German Subword Lexicon**
$\mathcal{LEX}_{EN} := \{$	$\mathcal{LEX}_{GE} := \{$
$(a, IV, \varepsilon, EN, d)$,	$(zunge, ST, \#tongue, GE, d)$,
$(hyoid, ST, \#hyoid, EN, d)$,	$(n, SF, \varepsilon, GE, d)$,
$(fracture, ST, \#fracture, EN, d)$,	$(bein, ST, \#bone, GE, d)$,
$(is, IV, \varepsilon, EN, d)$,	$(bruech, ST, \#bruch, GE, d)$,
$(rare, ST, \#rare, EN, d)$,	$(e, SF, \varepsilon, GE, d)$,
$(phenomenon, ST, \#phenomenon, EN, d)$,	$(sind, IV, \varepsilon, GE, d)$,
$(that, IV, \varepsilon, EN, d)$,	$(selten, ST, \#rare, GE, d)$,
$(may, IV, \#possible, EN, d)$,	$(ereignis, ST, \#phenomenon, GE, d)$,
$(result, ST, \#result, EN, d)$,	$(se, SF, \varepsilon, GE, d)$,
$(in, IV, \varepsilon, EN, d)$,	$(mit, IV, \varepsilon, GE, d)$,
$(signific, IV, \#signific, EN, d)$,	$(teils, IV, \#possible, GE, d)$,
$(ant, SF, \varepsilon, EN, d)$,	$(erheblich, ST, \#significant, GE, d)$,
$(complicat, ST, \#complic, EN, d)$,	$(en, SF, \varepsilon, GE, d)$,
$(ions, PS, \varepsilon, EN, d)$ $\}$	$(komplikat, ST, \#complic, GE, d)$,
	$(ionen, SF, \varepsilon, GE, d)$ $\}$

Subword Thesaurus
$\mathcal{THES}_d := (expandsTo, hasSense)$, with
$expandsTo := \{(\#hyoid, \#tongue), (\#hyoid, \#bone)\}$
$hasSense := \{(\#bruch, \#fracture), (\#bruch, \#hernia)\}$

Table 3.1: Example Lexicon for English, German and the Thesaurus

3.2.2 Subword Thesaurus

The subword thesaurus organizes equivalence classes of subwords, within and between different languages. Whenever lexical entries share the same MID and domain, they belong to the same equivalence class, or, the other way round, an equivalence class is defined by a subset of lexical entries: $\mathcal{C} \subset \mathcal{LEX}$. By convention, elements of this set are annotated with c, followed by the corresponding equivalence class symbol (MID) in subscript. For example, the set $c_{\#inflamm}$ contains all lexical items in different languages which have the meaning *inflammation*:

$$
\begin{aligned}
c_{\#inflamm} := \{\ & (inflamm, ST, \#inflamm, EN, d_1),\\
& (itic, SF, \#inflamm, EN, d_1),\\
& (itis, SF, \#inflamm, EN, d_1),\\
& (entzuend, ST, \#inflamm, GE, d_1),\\
& (itis, SF, \#inflamm, FR, d_1), \ldots\}
\end{aligned}
$$

Different MIDs can be linked by two lexical relations, *viz.* the horizontal (syntagmatic) relation $expandsTo \subset \mathcal{M} \times \mathcal{M}$, and the vertical (paradigmatic) relation $hasSense \subset \mathcal{M} \times \mathcal{M}$:

- The set $S_1 := \{(m_0, m_1), (m_0, m_2), \ldots, (m_0, m_n)\} \in expandsTo$ (with $m_{0,\ldots,n} \in \mathcal{M}$ and $|S_1| \geq 2$) relates a MID m_0 to a list of at least two MIDs. This relation is used in order to make a hidden semantic compositionality explicit. As an example, the MID assigned to the lexical item *short* is expanded to the MID representing the lexemes { *"length", "longitud", "compriment"*} and the MID representing the meaning of *small value.* The relation $expandsTo$ is also used to deal with composed meanings in compounds which exhibit omission of characters, e.g. *urinalysis* (see the discussion in the next chapter).

- The set $S_2 := \{(m_0, m_1), (m_0, m_2), \ldots, (m_0, m_n)\} \in hasSense$ (with $m_{0,\ldots,n} \in \mathcal{M}$ and $|S_2| \geq 2$) relates an ambiguous MID m_0 to a set of MIDs with at least two elements. It is used to link an ambiguous MID to each of its (non-ambiguous) senses. As an example, the MID assigned to the ambiguous word *head* is related via $hasSense$ to the non-ambiguous MIDs for *upper part of the body* and *person in charge of something.*

Both relations together constitute the thesaurus $\mathcal{THES}$ of a domain d:

$$\mathcal{THES}_d := (expandsTo, hasSense)$$

The sample thesaurus $\mathcal{THES}_d$ in Table 3.1 (bottom) consists of two elements for each of the relations. The word *"hyoid"* (also *"hyoid bone"* or *"tongue bone"*) can be translated to German *"Zungenbein"* (literally *"tongue bone"*[4]). It is derived from its

[4]Actually, the German stem *"bein"* has two senses: *"bone"* and *"leg"*. For simplicity, this ambiguity is not accounted for in the example.

anatomical location and, therefore, semantically composite. Consequently, *#hyoid* is expanded by the equivalence class symbols for both *"tongue"* and *"bone"*. For German, the word *"Bruch"*, which is assigned its own MID *#bruch*, is ambiguous and can be translated to either *"fracture"*, or *"hernia"* in English. These ambiguous readings are therefore coded in the relation *hasSense*.

Other than in many thesauri such as the UMLS (2005) or WordNet (Fellbaum, 1998), semantic relations between equivalence classes such as hypernymy, hyponymy, mereonymy etc. are not defined. Encoding these richer relations is left to domain thesauri or ontologies such as MeSH (2005) or CT (2004) to which lexical items can be mapped (Markó et al., 2003; Markó et al., 2004a; Hahn et al., 2004b; Markó et al., 2006c)

3.2.3 Subword Indexing

Subword lexicon and thesaurus are the declarative resources for the morpho-semantic normalization of (medical) texts. The third component, the subword indexer, constitutes the corresponding procedural component of the MorphoSaurus system. Input texts from languages under consideration are transcribed into a language-independent interlingua consisting of MIDs. It is based upon a three-step procedure (cf. Figure 3.2 for an English-German example based on the lexicons and thesaurus depicted in Table 3.1).

3.2.3.1 Orthographic Normalization

A preprocessor reduces all capitalized characters from input documents to lowercase characters and, additionally, performs language-specific character substitutions, (e.g., for German *'ß'* → *'ss'*, *'ä'* → *'ae'*, *'ö'* → *'oe'*, *'ü'* → *'ue'* and for Portuguese *'ç'* → *'c'*, *'ú'* → *'u'*, *'õ'* → *'o'*, cf. Figure 3.2, top-right). This eases the matching of (parts of) text tokens and entries in the lexicons. Additional translation rules are motivated by idiosyncrasies of the medical sublanguage, e.g. for German: *'ca'* → *'ka'*, *'co'* → *'ko'*, *'cu'* → *'ku'*, *'ce'* → *'ze'*, *'ci'* → *'zi'*, and others. This solves a notorious problem in German medical terminology (Brigl et al., 1994) where original Latin terms contain *'c'* instead of *'k'* and *'z'*, whereas German derivations of the same

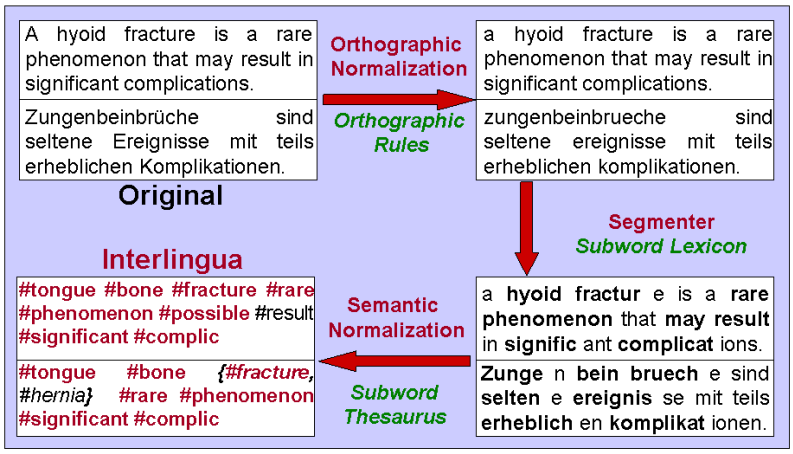

Figure 3.2: Subword Indexing Pipeline

terms prohibit the use of *'c'*, a rule frequently violated even by professional medical writers (e.g., the use of different surface forms such as *"Karzinom"*, *"Carzinom"*, *"Carcinom"* in German).

3.2.3.2 Morphological segmentation

The system segments the orthographically normalized input stream into a sequence of semantically plausible sublexical items, corresponding to subwords as found in the lexicon (cf. Figure 3.2, bottom right). The segmentation proceeds as follows: Each document token t of length n defined as a sequence of characters $c_1, c_2, \ldots, c_n$ is processed, in parallel, by a forward and backward matching process. The forward matching process starts at the positions 1 and $k = n$ and decrements k iteratively by one unless the sequence $c_1, c_2, \ldots, c_k$ is found in the subword lexicon. Alternatively, the backward matching process starts at the positions $k = 1$ and n and increments k iteratively by one unless the sequence $c_k, c_{k+1}, \ldots, c_n$ is found in the lexicon. Substrings recognized this way are entered into a chart. Unless the remaining sequences are not empty, $c_{k+1}, c_{k+2}, \ldots, c_n$, as well as $c_1, c_2, \ldots, c_{k-1}$ are tested recursively in the same manner, by forward and backward matching, respectively.

The segmentation results stored in the chart are checked for morphological plausibility using a finite-state automaton in order to reject invalid segmentations (e.g.,

segmentations without stems or beginning with a suffix, cf. Figure 3.1). If there are ambiguous valid readings or incomplete segmentations (due to missing entries in the lexicon), heuristic rules are applied, which prefer those segmentations with the longest match from the left and the lowest number of unspecified segments. Whenever the segmentation algorithm fails to detect a valid reading, all extracted stems of four characters or longer, if available, are preserved and the remaining fragments are discarded. Otherwise, if no stem longer than four characters can be determined during the segmentation, the original word is restituted. This method proved useful for the preservation of proper names.

3.2.3.3 Semantic Normalization

Each semantically relevant sublexical unit produced by the morphological segmentation is replaced by its corresponding MID which represents all subwords assigned to one particular class. If the MID has an entry in the thesaurus (i.e. there exist two or more MIDs related to it via $expandsTo$ or $hasSense$), that symbol is replaced accordingly. In the case of the $expandsTo$ relation, that particular MID is substituted by the sequence of related MIDs, e.g. $\#hyoid$ is exchanged by $\#tongue$ and $\#bone$ in the example in Figure 3.2 (bottom-left), based on the thesaurus in Table 3.1. For the $hasSense$ relationship, the ambiguity is marked with curly brackets and different readings are separated by commas (the German stem *"bruch"* is assigned the MID $\#bruch$, which is then replaced by $\{\#fracture, \#hernia\}$ in Figure 3.2). The result is a morpho-semantically normalized document in a language-independent, interlingual representation, where bold MIDs co-occur in both fragments. A comparison of the original natural language documents at the beginning of the pipeline and their interlingual representation at the very end already reveals the high degree of content similarity hidden by the natural language surface form.

Chapter 4

Implementation of the Subword Model

Within the MorphoSaurus system, the subword model was implemented covering the domain of clinical medicine for English, German, French, Spanish, Portuguese, and Swedish. The strategy for the creation, curation, and validation of the lexicon and thesaurus is described more detailed in the following.

4.1 Lexicon Creation

A comprehensive list of standard and domain-specific affixes is the starting point of subword lexicon building. Sources for affixes and infixes are the morphological grammar specification for the respective languages.[1] As a consequence, the main criterion for the delimitation of a word stem is its compatibility with existing prefixes and suffixes: *"in⊕compatib⊕ility"*, *"aprend⊕izaje"*, *"ventricul⊕i"*. Wherever derivation causes a clear change of the word sense which goes beyond the combined sense of the compounds, the derivate gains the status of a new lexeme with a different MID, e.g. *"decubit"* in addition to *"cubit"*, *"neurot"* in addition to *"neur"*. Many words of Latin and Greek origin come with stem variants (e.g., *"corpus"* vs. *"corpor⊕is"*,

[1] Common agglutination of suffixes may be pre-coded (e.g., *"-igkeiten"*, *"-izations"*, *"-ectomies"*, *"-ivelmente"*, *"-ingness"*, *"-ationally"*).

"abdomen" vs. *"abdomin⊕al"*, *"diagnos⊕is"* vs. *"diagnost⊕ico"*). Here, a reduction to the common substring (*"corp"* or *"abdom"*) would cause the proliferation of pseudo-suffixes (here *"-oris"*, *"-inal"*) on the one hand and the generation of short word stems on the other hand. In these cases stem variants are added to the lexicon as synonyms.

4.1.1 Delimiting Subwords

A high-performance extraction of subwords from large amounts of text is achieved by the use of finite-state techniques for lexicon-based decomposition, dederivation and deflection such as described above. Lexicon builders' decisions of subword delimitation are therefore driven not only by formal linguistic criteria, but also by the proper function of segmentation using finite-state machines. This is especially relevant to long and composed words where different valid segmentations are possible. For example, using a subword lexicon for English in domain d, *"nephrotomy"* may be segmented into the sequence of lexical units

$$(nephr, ST, \#kidney, EN, d)\oplus$$
$$(o, IN, \varepsilon, EN, d)\oplus$$
$$(tomy, PS, \#incision, EN, d)$$

but also into

$$(nephr, ST, \#kidney, EN, d)\oplus$$
$$(oto, ST, \#ear, EN, d)\oplus$$
$$(my, ST, \#muscle, EN, d)$$

If the word segmentation routine, here, prefers a long match from the left, the second (erroneous) segmentation is preferred. Only costly knowledge and language processing routines (which are not available, in general) would be expected to detect this kind of errors. A pragmatic solution is to include additional synonymous lexeme variants. In the example, this means that

$$l_{13} = (nephro, ST, \#kidney, EN, d)$$

is added to the English lexicon, and correspondingly,

$$l_{14} = (nefro, ST, \#kidney, SP, d)$$
$$l_{15} = (nefro, ST, \#kidney, PT, d)$$

to the Spanish and Portuguese one.

4.1.2 Empirical Validation of Subword Specificity

Especially short or ambiguous word stems, such as *"gen", "my", "mi", "ship"* are prone to side effects as described above. The shorter they are, the more frequently they occur as accidental substrings, producing erroneous segmentation results. In order to empirically assess this risk, lexical entries are matched against word lists derived from domain-specific text corpora. Two cases can then be distinguished:

The number of accidental matches is high: First, all correct matches have to be checked. Here, in many cases, the short stem will occur at the beginning of a word. If this does not lead to false matches, this stem can be (unorthodoxly) added as a proper prefix (PP) in order to make use of the position constraint on this class of lexemes. If there are still many occurrences in the inside of words left, then, the pertaining compounds or prefix-stem combinations have to be added to the lexicon and linked to their components by expansion. An example is the stem *"ship"*: It has to be avoided that the sense of *"ship"* (vessel, to send) is extracted from any word with the suffix *"-ship"*, e.g. *"relationship"*. Therefore, instead of defining a stem, *"ship"* is added as an invariant, as well as a (purely formal) prefix (*"ship⊕men"*):

$$l_{16} = (ship, SF, \varepsilon, EN, d)$$
$$l_{17} = (ship, IV, \#ship, EN, d)$$
$$l_{18} = (ship, PF, \#ship, EN, d)$$

Moreover, inflectional forms and derivatives of short verbs have to be included in the lexicon as invariants, e.g. for the MID $\#eat$:

$$l_{19} = (eat, IV, \#eat, EN, d)$$
$$l_{20} = (eats, IV, \#eat, EN, d)$$

$l_{21} = (eating, IV, \#eat, EN, d)$
$l_{22} = (ate, IV, \#eat, EN, d)$
$l_{23} = (eaten, IV, \#eat, EN, d)$
$l_{24} = (eater, IV, \#eat, EN, d)$

There are relatively few accidental matches: Here, the strategy is the opposite one. The stem is added to the lexicon, and the erroneously matching words are segmented. Wherever the erroneous stem happens to be extracted, adjustments have to be made to the components of these words. An example for this is *"oto"* in the word *"nephrotomy"* (see discussion above). Instead of eliminating *"oto"* as a stem, the stem variant *"nephro"* is added to the lexicon and, thus, false segmentation results are avoided.

4.1.3 Criteria for Lexical Subword Inclusion

The selection of lexical units should reflect the language use in the domain of interest. Again, word statistics extracted from extensive, language-specific corpora are used in order to measure the relevance of terms. Ideally, each lexicon entry should correspond to an atomic (indivisible) entity of semantic reference. However, there are borderline cases, especially where a composed lexeme may have an atomic synonym. As consequences, either the composed lexeme is entered as a whole (as a multi-word term) and equalized with its atomic synonym, or the atomic lexeme is related to the components of its synonym by the relation *expandsTo*. For example, the English adjective *"ascorb⊕ic"* implies *"vitamin c"* (other languages accordingly):

1. $l_{25} = (ascorb, ST, \#ascorb, EN, d)$
   $l_{26} = (vitamin\ c, IV, \#ascorb, EN, d)$
2. $\{(\#ascorb, \#vitamin), (\#ascorb, \#c)\} \in expandsTo$

The latter case is preferred if the components of the composed lexeme are semantically relevant. But in this example, the first one is favored since the MID $\#c$ is semantically weak.

In contrast to the general rule, semantically underdetermined complex lexemes or noun groups need not to be included in the dictionary as long as there exists

a strict mapping through all languages of interest. As an example, the sense of the term *"yellow fever"* is not derivable from its components, but its components literally translate to all languages (e.g. Spanish *"fiebre amarilla"*, Portuguese *"febre amarela"*, or German *"Gelbfieber"*).

Proper names are entered into the lexicon under the following circumstances:

1. They are needed for synonym linkage, e.g. between different product names, e.g.
   $l_{27} = (diclofenac, IV, \#diclofenac, EN, d)$
   $l_{28} = (voltaren, IV, \#diclofenac, EN, d)$
   $l_{29} = (cataflam, IV, \#diclofenac, EN, d)$

2. They are used as eponyms, i.e. they belong to the domain terminology, e.g.
   $l_{30} = (crohn, IV, \#crohn, EN, d)$
   $l_{31} = (parkinson, IV, \#parkinson, EN, d)$

3. Translations exist, especially with regard to geographic terms, e.g.
   $l_{32} = (switzerland, IV, \#switzerland, EN, d)$
   $l_{33} = (suisse, IV, \#switzerland, FR, d)$

4.2 Thesaurus Creation

As introduced above, equivalence class identifiers can be linked using the semantic relations *hasSense* and *expandsTo*. Groups of lexemes which have (the same) multiple senses are assigned a MID of their own. The *hasSense* relation then connects such ambiguous MIDs to each of their senses. For example, the Spanish word *"lobo"* is ambiguous, since it may refer to an animal ($\#wolf$), or to an anatomical object ($\#lobe$). Therefore, for the lexical entries

$l_{34} = (lobo, IV, \#lobo, SP, d)$
$l_{35} = (wolf, ST, \#wolf, EN, d)$
$l_{36} = (wolves, ST, \#wolf, EN, d)$
$l_{37} = (lob, ST, \#lobe, EN, d)$

the following thesaurus relation is added:

$\{(\#lobo, \#wolf), (\#lobo, \#lobe)\} \in hasSense$

The *expandsTo* relation links one or more non-atomic lexemes (which are also grouped by a separate MID) to their atomic senses. There are mainly four reasons for this:

1. Utterly short morphemes are not permitted as word constituents in order to prevent improper segmentation of compounds. Words which contain these morphemes must therefore have their semantic decomposition pre-coded. For example, for the entries
   $l_{38} = (myalg, ST, \#myalg, EN, d)$
   $l_{39} = (mialg, ST, \#myalg, SP, d)$
   $l_{40} = (muscle, ST, \#muscle, EN, d)$
   $l_{41} = (muscul, ST, \#muscle, SP, d)$
   $l_{42} = (pain, ST, \#pain, EN, d)$
   $l_{43} = (algia, SF, \#pain, SP, d)$
   the relation *expandsTo* is extended by:
   $\{(\#myalg, \#muscle), (\#myalg, \#pain)\} \in expandsTo$
   thus, avoiding the occurrence of *"my"* or *"mi"* in the lexicon.

2. A non-decomposable lexeme in one language has a composed sense in the reference language (English). For example:
   $l_{44} = (esparadrap, ST, \#esparadrap, SP, d)$ and
   $\{(\#esparadrap, \#adhesiv), (\#esparadrap, \#tape)\} \in expandsTo$

3. Compounds exhibit ellipsis (omission of characters). For example:
   $l_{45} = (urinalise, ST, \#urinalise, PT, d)$ and
   $\{(\#urinalise, \#urin), (\#urinalise, \#analys)\} \in expandsTo$

4.3 Aspects of lexicon construction

The delimitation of classes of semantic equivalence is mainly an intellectual task which cannot be fully automatized. As a starting point, each lexicon entry has its

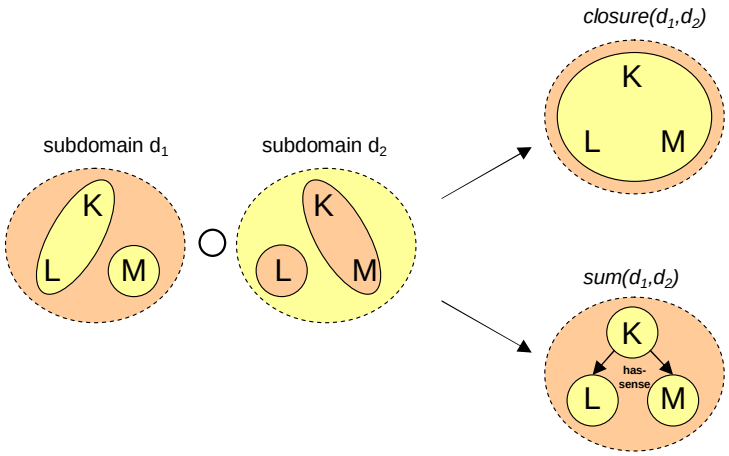

Figure 4.1: Fusing Subdomains

own MID. If the lexicon designer concludes that two lexicon entries have an identical sense, then the two MIDs are fused.

The incremental fusion of lexemes, however, repeatedly leads to a class of decisions which can be considered as the main dilemma of the lexicon engineering process. Fig. 4.1 illuminates this situation. Let K, L, and M be atomic lexical items. Two lexicographers may group these items in different ways, according to slightly different subdomain contexts, here represented by d_1 and d_2, respectively. In d_1 the lexical items K and L are considered synonyms. In d_2, however, M instead of L is considered a synonym of K. The fusion of these two subcontexts gives two solutions, *viz.* closure and sum, as depicted in Fig. 4.1 (right). Whereas the closure operation merges the synonym classes, the sum operation preserves the context-related distinction and introduces two senses for the ambiguous equivalence class. The reasons for the decision whether one follows the one or the other strategy are quite complex. On the one hand, a tight network of ambiguous senses results from pursuing the latter strategy. On the other hand, the transitive closure tends to yield numerous synonym classes in which pairs of lexemes may hardly be synonymous anymore. As an example, a user may assert synonymy between *"head"* and *"caput"* in an anatomy subdomain. Another one equalizes *"head"* with *"chief"*, when modeling terms in a subdomain of administration. Applying the closure operation, *"chief"* would become synonym to *"caput"*, and all literal and figurative

senses of *"head"* would be represented by one MID. Applying the sum operation, *"head"* would be assigned an ambiguous MID which then would be related to its non-ambiguous senses.

4.3.1 A Web-based Lexicon Editing Tool

A powerful editor for subword lexicons was developed to facilitate the work of the lexicographers. The tool is Web-based, so that different users at different places (speaking different native languages) can work on the same lexical repository. Its interface is tiled vertically and consists of two identical windows which allow to easily browse through the lexicon, join two different equivalence classes or link them using the *expandsTo* or *hasSense* relation. It allows numerous different sorting and constraint criteria when browsing through the lexicon. In addition, it offers different word statistics features. For example, a lexicon curator may have a look at the word frequencies of large domain-specific word lists containing a particular substring. This proved to be useful in determining whether a short word stem should be integrated into the subword lexicon, or not. Figure 4.2 shows a screenshot of the lexicon editor.

4.3.2 Lexicon Statistics

Early investigations of the subword approach already revealed one of its benefit (Schulz & Hahn, 2000). For covering a particular domain (diagnosis reports), instead of spelling out derivational and compositional forms of medical terms which would increase the size of underlying lexicons dramatically by the sheer number of different term variants, Schulz & Hahn (2000) found a convenient growth behavior as far as the number of subword entries required are concerned. While incrementally accumulating a subword lexicon by stepwisely analyzing a corpus consisting of 30,000 diagnosis phrases, the lexicon growth they observed can be approximated by a well-known logarithmic function. For covering the whole corpus, a comparatively small list of 4,098 word stems were obtained.

Since this study, the lexical resources were manually constructed over the last five years with a changing amount of manpower. The English, German and Portuguese

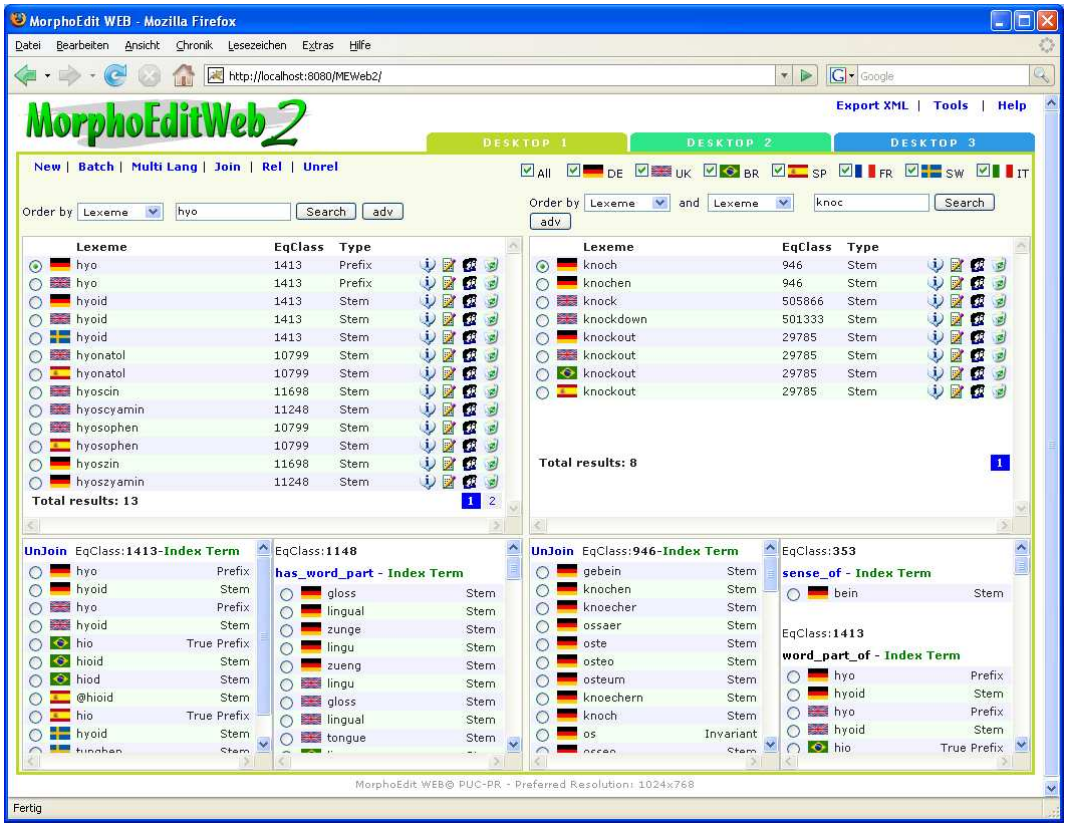

Figure 4.2: MorphoEdit Web

subword lexicons were created in a fully manual fashion, while for Spanish, French and Swedish, machine learning techniques were applied prior to manual work in order to stepwisely augment the lexicons.

Table 4.1 contains the most important data of the lexicons. Overall, the lexical resources contain 90,550 entries (Column 2),[2] from which 87,439 (Column 4) are linked to a total of 21,432 equivalence classes (Column 5). Thus, the lexicons contain 3,111 stop entries.

With more than 22,500 entries each, the English and German lexicons provide the highest coverage, followed by the Portuguese lexicon with about 15,000 entries. In

[2]Just for comparison, the size of WordNet (Fellbaum, 1998) assembling the *lexemes* of general English in the 2.0 version is on the order of 152,000 entries (`http://www.cogsci.princeton.edu/~wn/`). Linguistically speaking, the entries are basic forms of verbs, nouns, adjectives and adverbs.

Language	Subwords	Learned	Linked	EqClasses	Ratio
English	22,561	-	22,067	16,148	1.37
German	23,976	-	23,225	15,978	1.45
Portuguese	14,984	-	14,170	9,886	1.43
Spanish	10,936	8,793	10,387	7,408	1.40
French	7,812	5,777	7,556	5,351	1.41
Swedish	10,281	7,470	10,034	6,003	1.67
All	90,550	22,040	87,439	21,432	4.08

Table 4.1: Number of Subwords and their Linkage to the Thesaurus

contradistinction to this fully manual work, a total of 22,040 subwords were acquired automatically for Spanish, French and Swedish (Column 3, cf. next Chapter). There are between 1.4 to 1.7 subwords linked to one particular equivalence class within one language (synonymy), and 4.1 across the six different languages (synonymy and translation). In terms of relations between equivalence classes, there are currently 953 distinct *expandsTo* and 2,612 *hasSense* relations defined in the thesaurus.

As a particular benefit, the subword approach reduces the number of types needed to sufficiently cover a certain domain. Instead of collecting all derivational and compositional forms of medical terms which would cause the size of underlying lexicons to grow dramatically, the amount of subwords remains manageable.

In order to express this implicit assumption in figures, a subset of MEDLINE abstracts was built, some of which with full text reference to the corresponding German article. For English, the corpus was comprised of more than 155 million tokens, while the German collection contained 23 million words. Subsequently, the number of types required to cover certain percentages of these corpora was determined by counting the additive frequencies of the corresponding tokens within these documents (beginning with the most frequent ones) and by dividing these values by the total number of tokens in the medical corpora. Figures 4.3 and 4.4 show the typical asymptotic behavior of such curves (see line *"Unique Words"*).

After transforming the medical corpora into the interlingua, the same procedure was performed on these corpora. Again, the coverage of the collections in terms of

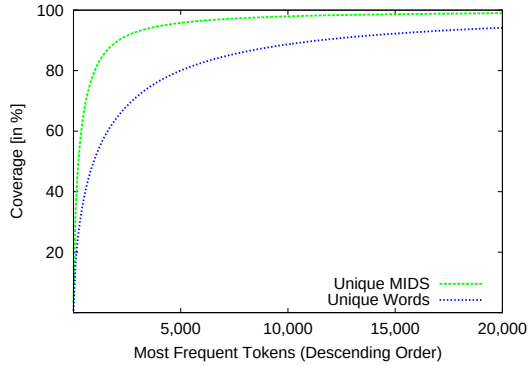

Figure 4.3: Coverage for English

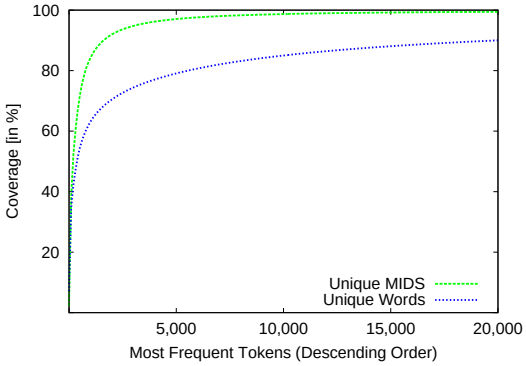

Figure 4.4: Coverage for German

Words	Coverage	Full Forms	Subwords
English			
124,469,156	80%	5,000	1,000
139,973,802	90%	11,500	2,100
147,750,124	95%	22,500	4,300
148,986,220	97%	27,200	5,000
153,971,182	99%	97,300	19,900
155,526,447	100%	528,587	276,846
German			
18,404,098	79%	5,000	800
20,943,138	90%	20,000	1,700
22,106,646	95%	53,400	3,200
22,586,337	97%	100,400	5,000
23,037,452	99%	248,600	12,400
23,270,154	100%	476,911	82,124

Table 4.2: Number of Entries to Cover English and German Medical Terminology

considering the most frequent words was taken into account (*"Unique MIDs"*). For both scenarios, Table 4.2 shows the corresponding number of tokens that are needed to cover certain percentages of the corpora.

As the figures show, the usage of subwords remarkably reduces the number of types (lexical entries) that are needed to cover these corpora. This, however, may sometimes be accompanied by a subtle loss of semantic distance that arises from joining quasi synonymous subwords (such as *"belly"* and *"abdomen"*) in a common equivalence class. For English, 22,500 full forms cover 95% of the collection, while there are only 4,300 subwords needed. To cover 99% of the corpora, nearly 100,000 full forms are required, compared to 20,000 lexical entries at the subword level. This effect is even more striking for German with a high amount of (ad hoc) nominal compounds. Only 12,400 subword entries are sufficient to cover 99% of the test collection. On the other hand, nearly 250,000 full forms are required for the same scenario.

The project originally started from a bilingual German-English lexicon, while the Portuguese part was added in a later project phase. Of course, the manual creation and maintenance of lexicons and the thesaurus in which equivalence classes are organized is costly and error-prone. In an effort to further expand the language coverage of the MorphoSaurus system by Spanish, French and Swedish, already available resources for Portuguese, English, and German are reused in order to speed up and to ease the lexicon acquisition process (Schulz et al., 2004; Markó et al., 2005a; 2005d; 2005f; 2006d).

Chapter 5

Lexical Acquisition

The bottleneck for dictionary-based natural language processing systems is the lack of comprehensive dictionaries, especially for many different languages in particular domains. In the following, a methodology is introduced by which multilingual subword dictionaries for Spanish, French and Swedish emerge automatically from simple seed lexicons. The creation of the initial lexicons for the languages in focus relies on cognate mapping, i.e., string-pattern-based transformations of orthographically very similar lexical forms from the source language into the target language. This seed is then thrown onto parallel corpora in order to filter out valid lexical and semantic hypotheses. For this step, the focus lies on co-occurrence patterns of hypothesized translation equivalents in the parallel corpora. Subsequently, valid cognates contribute to further dictionary upgrades by iteratively incorporating non-cognates into the lexical assimilation process.

5.1 Cognate Mapping

It is well known that typologically related languages reveal similarities both at the lexical and grammatical level. With these considerations in mind, it is an obvious idea to reuse already available resources from given languages to build up corresponding resources for other typologically related ones. The language pairs considered here are Portuguese/Spanish, English/French, German/French, English/Swedish as well as German/Swedish.

44 Rules:	**Portuguese**	**Spanish**	**18 Rules:**	**English**	**French**
ss → s	fracass	fracas	o → ou	movement	mouvement
lh → j	mulher	mujer	ve → f	nerve	nerf
+ça → za	cabeça	cabeza	+or → eur	receptor	recepteur
19 Rules:	**German**	**Swedish**	**26 Rules:**	**German**	**French**
ei → e	bein	ben	v → f	intensiv	intensif
+aa+ → a	saal	sal	s → z	gas	gaz
+u+ → ö	brust	bröst	or → eur	tumor	tumeur
7 Rules:	**English**	**Swedish**			
c → k	cramp	kramp			
ph → f	phosphor	fosfor			
ce → s	iceland	island			

Table 5.1: Some String Substitution Rules and Examples

From the Portuguese (alternatively, English and German) dictionary, identical and similarly spelled Spanish (French and Swedish) subword candidates are generated. As an example, the Portuguese word stem *"estomag"* (*"stomach"*) is identical with its Spanish cognate, while *"mulher"* (Portuguese, in English *"woman"*) is similar to *"mujer"* (Spanish). Similar subword candidates are generated by applying a set of string substitution rules, some of which are listed in Table 5.1. In total, 44 rules for Portuguese-Spanish were formulated, 18 rules for English-French, 19 rules for German-Swedish, 26 rules for German-French, and 7 rules for English-Swedish. These rules were all formulated by medical linguists based on introspection, also using various dictionaries for heuristic guidance. Some of these substitution patterns cannot be applied to starting or ending sequences of characters in the source subword. This constraint is captured by a wildcard ('+' in Table 5.1), which stands for at least one arbitrary character.

Based on these string substitution rules and the already available (Portuguese, English, German) lexicons, for each entry (excluding affixes) of these sources, all possible Spanish, French and Swedish variant strings were generated. This led,

Language Pair	String Variants			
	#Variants	4-chars	16-chars	overall
Portuguese-Spanish	123,385	2.7	89.8	8.8
English-French	47,020	1.6	5.3	2.2
German-French	74,994	2.0	35.4	3.3
English-Swedish	70,178	1.8	9.9	3.2
German-Swedish	152,819	2.6	37	6.7

Table 5.2: Variant Generation Statistics

on the average, to 8.8 Spanish variants per Portuguese subword (ranging from 2.7 for high-frequent four-character words to 89.8 for low-frequent 16-character words). Since the rule set is much smaller for the other language pairs, their average is far less than for Portuguese-Spanish, as shown in Table 5.2: For each language pair (first column), the total number of variants is depicted in the second column. Columns three to five show variant averages per length.

5.1.1 Cognate Candidate Elimination

All generated Spanish, French, and Swedish variants were subsequently compared with word frequency lists for these target languages which were compiled from large, heterogeneous medicine-related Web sources.

5.1.1.1 Resources

Corpus sources (a total of 2 GB) for all languages considered and their statistics are depicted in Table 5.3. They were derived from MEDLINE (English) or MEDLINE-related databases (other languages), i.e. abstracts of scientific publications in a particular language that are linked from MEDLINE to their original, language-specific source.[1] The contents of different medicine-related Web portals addressing physi-

[1] E.g., `http://www.springerlink.com/`

Language	**Corpus Type**	**Tokens**	**Types**
English	MEDLINE	209,302,337	
	Others	40,938,064	
	$\sum$	250,240,401	528,585
German	MEDLINE	1,327,435	
	Others	29,518,426	
	$\sum$	30,845,861	467,909
Portuguese	MEDLINE	200,446	
	Others	13,704,344	
	$\sum$	13,904,790	138,248
Spanish	MEDLINE	357,532	
	Others	11,103,066	
	$\sum$	11,460,598	126,314
French	MEDLINE	1,810,567	
	Others	2,355,541	
	$\sum$	4,166,108	85,710
Swedish	MEDLINE	—	
	Others	2,480,573	47,343
$\sum$		319,968,430	1,216,325

Table 5.3: Corpus Resources

cians and health care consumers served as additional resources.[2] Due to unbalanced availability (especially with regard to MEDLINE abstracts) the corpora obtained varied significantly. For English, a total of one million MEDLINE abstracts were included in the corpora, for German 8,000, for French 9,900, for Portuguese 1,370, and for Spanish 1,441. Unfortunately, to the best of knowledge, there are no MEDLINE abstracts which link to a Swedish source. The resources depicted here have been used in other experiments of this work, as well (cf. Chapters 6, 7, 9 and 10).

[2]E.g., different language versions of *Netdoctor*, cf. `http://www.netdoctor.co.uk/` and the Merck Manual of Diagnosis and Therapy, cf. `http://www.msd.de/msdmanual/home.html`

5.1.1.2 Elimination of Cognate Candidates

Wherever a (purely formal) prefix string match (in the case of stems) or an exact match (in the case of invariants) occurred in the generated corpora, the matching string was listed as a potential target cognate of the source language subword it originated from. Whenever several substitution alternatives for a source subword had to be considered, that particular alternative was chosen which had the most similar lexical distribution in the corpora considered. Similarity was measured as follows: Let S be the source lexical item, C_S the source language corpus containing n tokens and V_1, V_2, ..., V_p the hypotheses generated from S that match the target language corpus C_T, containing m tokens. With $f(x, y)$ denoting the frequency of a word x in a corpus y, that particular V_j $(1 \leq j \leq p)$ was chosen for which

$$\left| \frac{f(S, C_S)}{n} - \frac{f(V_j, C_T)}{m} \right|$$

was minimal. All other candidates were discarded.

As a result, a list of putative target language subwords was obtained, each one linked by the associated MID to their grounding cognate in the source lexicon. These lists of cognate candidates are referred to CC_{SPA} for Spanish, CC_{FRE} for French, and CC_{SWE} for Swedish.

Starting from 14,114 Portuguese, 23,259 German and 22,014 English subwords (only considering stems and invariants), a total of 123,385 Spanish subword variants were created using the string substitution rules. For Swedish (French), 152,819 (74,994) variants were derived from German and 70,178 (47,020) from English (cf. Table 5.2). Matching these variants against the Spanish corpus and allowing for a maximum of one candidate per source subword, 11,161 tentative Spanish cognates were identified. Combining English and German evidence, 11,930 French and 7,024 tentative Swedish cognates were found (cf. Table 5.4). Spanish candidates were linked to a total of 8,219 MIDs from their Portuguese correlates (hence, 2,942 synonym relationships have also been hypothesized), whilst French (Swedish) candidates were associated with 8,218 (4,634) MIDs from their German and English correlates.

Language Pair	Source Lexicon	Selected Cognates	Linked MIDs
Portuguese-Spanish	14,114	11,161	8,219
English-French	22,014	9,672	7,373
German-French	23,259	8,551	6,737
Combined Evidence		11,930	8,218
English-Swedish	22,014	4,512	3,440
German-Swedish	23,259	4,982	3,740
Combined Evidence		7,024	4,634

Table 5.4: Selected Cognates (Including Combined Evidence for French and Swedish)

5.2 Cognate Validation Using Parallel Corpora

Large multilingual resources which are available in the biomedical domain were used in order to identify false cognates (so-called *false friends*, i.e., similar words in different languages with different meanings. For example, the Spanish subword candidate **"crianz"* for the Portuguese *"crianc" ["child"]* (the normalized stem of *"criança"*) was found in the list of generated cognate-pairs. The correct translation of Portuguese *"crianc"* to Spanish, however, would have been *"nin"* (the stem of *"niño"*), whilst the Spanish *"crianz"* refers to *"criac" ["breed"]* (stem of *"criação"* in Portuguese).

The corpus used here was derived from the *Unified Medical Language System* UMLS (2005)[3], a collection of different medical terminology systems, such as the *International Classification of Diseases* (ICD-10, 2005) or the *Medical Subject Headings* (MeSH, 2005) (cf. the end of Section 3.1).

Entries of these different resources are linked to each other via the UMLS Metathesaurus, which makes it possible to extract translations of terms for various languages. Unfortunately, word-to-word translation occurs only in very few

[3]See Section 12.2 for a selection of non-medical resources (covering policy, law, economics, culture, education, etc.), which can be used just in the same manner.

cases. More often one encounters rather complex noun phrases with a similarly complex semantic structure. Examples for typical English-Spanish alignments are *"Cell Growth"* aligned with *"Crecimiento Celular"*, or *"Heart transplant, with or without recipient cardiectomy"* aligned with *"Trasplante cardiaco, con o sin cardiectomia en el receptor"*, which reveal a phrasal level of semantic correspondence.

English was used as the pivot language for the validation of generated cognates, since it has the broadest coverage in the UMLS. The linkage to other languages is considerably poorer, both in qualitative as well as quantitative terms. The size of the corpora derived from the linkages of the English UMLS to other languages amounts to 60,526 term translations for English-Spanish,[4] 17,130 for English-French, and 10,953 alignments for English-Swedish. Furthermore, additional 28,473 English-Swedish alignments were made available by Nyström et al. (2006), thus summing up to 39,426.

In order to determine the false cognates in the lists of the generated cognate pairs, CC_{SPA}, CC_{FRE} and CC_{SWE}, these lists served as preliminary lexicons for the morpho-semantic normalizer, including 836 manually added affixes for Spanish, 279 for French, and 601 for Swedish. Based on these subword resources, the parallel corpora of the aligned UMLS expressions were then morpho-semantically processed as described in the previous chapter.

Whenever the same MID occurred on both sides after this simultaneous bilingual processing of the UMLS alignments, the appropriate Spanish (French or Swedish, alternatively) subword entry that led to this particular MID was taken to be a valid entry. This is a reasonable approach, since it is highly unlikely that a false friend occurs within the same translation context.

Those hypotheses which never matched in this validation procedure were rejected from the candidate lexicons. As a result (cf. Table 5.5), 49% of the Spanish, 33% of the French, as well as 34% of the Swedish hypotheses were acknowledged. Together with the manually provided list of affixes, the list of accepted cognates served as the seed lexicons (in the following, $\mathcal{L}(0)$) for acquiring additional lexical entries, which were *not* cognates to elements of any of the source lexicons.

[4]Only focusing on the so-called *preferred entries*.

Language Pair	**Hypotheses**	**Valid**	$\mathcal{L}(0)$ **incl. Affixes**
Portuguese-Spanish	11,161	5,481 (49.1%)	6,317
English/German-French	11,930	3,903 (32.7%)	4,182
English/German-Swedish	7,024	2,384 (33.9%)	2,985

Table 5.5: Cognates Matching the UMLS Alignments

5.3 Bootstrapping Subwords

The parallel corpora derived from the UMLS and the lexicons with validated cognates both served as starting points for a continuation of the lexical acquisition process, as described in Algorithm 1. In order to illustrate this process, assume the Swedish subword *"blod"* was identified as being a cognate to the English subword *"blood"* (and, therefore, is included in $\mathcal{L}(0)$). Then, the yet unknown Swedish word *"blodtryck"*, which has the English translation *"blood pressure"* in the UMLS Metathesaurus gets segmented into

$(blod, ST, \#blood, SW, d)\oplus$
$(t, UK, \varepsilon, SW, d)\oplus$
$(r, SF, \varepsilon, SW, d)\oplus$
$(yck, UK, \varepsilon, SW, d)$

with ST being a marker for a stem, SF for a suffix and UK for an unknown sequence for Swedish (SW) in domain d, thus satisfying the condition in line 12 of the algorithm. At the same time, the morpho-semantic normalization of *"blood pressure"* leads to the sequence of MIDs [*#blood #tense*], whilst the normalization of *"blodtryck"* leads to [*#blood*], since *"tryck"* is not yet part of the Swedish lexicon. Comparing these two representations, the condition in line 13 of the algorithm is satisfied, since there is exactly one more MID resulting from the English decomposition which cannot be found in the Swedish normalization result. The invalid segment is then reconstructed (*'t⊕r⊕yck'*) by eliminating those substrings that led to a matching MID (*"blod"*) in the aligned unit (*"blodtryck"*) (line 15). The supernumerary MID resulting from the English normalization is assigned to that remaining

1: MSI: morpho-semantic indexing procedure described in Section 3.2 (maps sequences of words to sequences of MIDs and remainders)
2: $current \leftarrow 0$
3: $quiescence \leftarrow$ false
4: **while** not $quiescence$ **do**
5: the lexicon for MSI is set to $\mathcal{L}(current)$
6: the list of $new\_entries$ is empty
7: **for all** AU_i, $i \in$ [1,n] (UMLS alignment units) **do**
8: $AU_S \leftarrow$ source language part of AU_i
9: $AU_T \leftarrow$ target language part of AU_i
10: $MID_S \leftarrow MSI(AU_S)$
11: $MID_T \leftarrow MSI(AU_T)$
12: **if** for exactly one word there is an invalid segmentation (checked by the FSA) in MID_T **then**
13: **if** there is exactly one more MID in MID_S than in MID_T **then**
14: $mid \leftarrow$ supernumerary MID from MID_S
15: $entry \leftarrow$ restore the invalid segment and remove substrings that led to a matching MID in MID_S and MID_T;
16: strip off potential suffixes from $entry$, if the remaining substring is longer than 4 (thus, avoiding too short entries);
17: add $entry$ together with the associated mid to $new\_entries$
18: **end if**
19: **end if**
20: **end for**
21: **if** $new\_entries$ is empty **then**
22: $quiescence \leftarrow$ true
23: **else**
24: $current \leftarrow current + 1$
25: copy $\mathcal{L}(current - 1)$ to $\mathcal{L}(current)$
26: add all entries from $new\_entries$ to the lexicon $\mathcal{L}(current)$
27: **end if**
28: **end while**

Algorithm 1: Bootstrapping Algorithm for Lexical Acquisition

Lexicon	**Spanish**	**French**	**Swedish**
$\mathcal{L}(0)$	6,317	4,182	2,985
$\mathcal{L}(1)$	8,610 (2,293)	5,679 (1,497)	6,893 (3,908)
$\mathcal{L}(2)$	8,771 (161)	5,768 (89)	7,347 (454)
$\mathcal{L}(3)$	8,788 (17)	5,777 (9)	7,388 (41)
$\mathcal{L}(4)$	8,793 (5)	5,777 (0)	7,467 (79)
$\mathcal{L}(5)$	8,793 (0)		7,470 (3)
$\mathcal{L}(6)$			7,470 (0)
$\mathcal{L}_{ALL}$	**10,936** (2,143)	**7,812** (2,035)	**10,281** (2,811)

Table 5.6: Lexicon Growth Steps (Δ in brackets)

substring (line 17 in the algorithm). After processing all UMLS alignments, this new entry is then incorporated in the Swedish lexicon as a stem, resulting in the lexicon $\mathcal{L}(1)$ (line 26). In the next run, in which all UMLS alignments are processed once again, this newly derived lexicon entry may serve for extracting, e.g., the Swedish word *"luft"* with its identifier *#aero* from the UMLS entry *"air pressure"* (English, indexed to [*#aero #tense*]) linked to *"lufttryck"* (Swedish). When no new entries can be generated using this method (quiescence), the algorithm stops.

Table 5.6 depicts the growth steps of the target lexicons for the entire bootstrapping process (new entries in comparison to each previous step are in brackets). In the first run, for Spanish, 2,293 new lexemes were added to the lexicon which comes to a size of 8,610 including those lexemes already generated by the cognate identification routines (cf. Table 5.5). For French, 1,497 new lexemes were generated in the first step and for Swedish 3,908. After four runs, learning came to an end with 8,793 lexemes generated for Spanish, while after three runs, 5,777 lexicon entries for French were acquired. Finally, for Swedish, 7,470 lexemes were learned after five iteration steps.

These automatically acquired lexicons served as the basis for the additional manual enhancements of the lexicons involved. In the meantime, 2,143 Spanish, 2,035 French and 2,811 Swedish subwords have been added by hand, resulting in a total

of 10,936 entries for Spanish, 7,812 for French and 10,281 for Swedish, referred by $\mathcal{L}_{ALL}$ in Table 5.6.

5.4 Checking the Quality of Derived Lexicons

For lexicon generation, Portuguese-Spanish, English/German-French, and English/German-Swedish corpora compiled out of the UMLS Metathesaurus were used. To estimate the quality of the interlingual connections between the newly derived lexicons, the results after running the morpho-semantic indexing system (the function *MSI* from Algorithm 1, as described in Section 3.2) on these collections were compared, at each stage of the lexical acquisition process. Of course, these results probably include overfitting phenomena. Therefore, additional Spanish-French (13,158), Spanish-Swedish (8,993) and French-Swedish (6,713) aligned entities were extracted from the UMLS. The alignments range, again, from word-to-word translations (e.g., Spanish *"pierna"* to Swedish *"ben"* (English *"leg"*)) to complex noun phrases, which sometimes correspond to a single word in the other language, e.g., the Spanish phrase *"enfermedad vírica transmitida por artrópodos, no especificada"* maps to the Swedish *"arbovirusinfektioner"* (English *"arbovirus infections"*) in the UMLS. This example also reveals the problem of inexact translations (especially for data coming from the *International Classification of Diseases* (ICD-10, 2005) and the *International Classification of Primary Care* (ICPC, 1990)). The Spanish fragment *"no especificada"* (*"not specified"*), e.g., does not have a counterpart in the Swedish equivalent.

Rather than only examining the coverage of the acquired lexicons, the quality of the generated lexicons (admitting that their status is far from being complete[5]) was estimated, i.e. the validity of the interlingual synonymy relations stipulated.

For this goal, the English-Spanish, English-French, and English-Swedish corpora, on which the lexical acquisition was based employing the MSI routines for all lex-

[5]With lexicon sizes from 7,812 for French, 10,281 for Swedish to 10,936 entries for Spanish the lexicons are certainly far from being complete (compared to 22,561 entries for English, 23,976 for German and 14,984 for Portuguese, see Section 4.3.2).

icon levels, $\mathcal{L}(0)$- $\mathcal{L}(5)$, were indexed. Furthermore, the Spanish-Swedish, Spanish-French, and French-Swedish corpora (previously unseen by the learning algorithm) were processed accordingly. For each alignment unit of the corpora, the resulting MIDs were then compared using a measure of indexing consistency proposed by Hooper (1965):

$$C_{AU_i} = \frac{100A}{A + N + M}$$

The indexing consistency of one alignment unit (AU_i) of the parallel corpus, C_{AU_i}, is dependent on A, the number of MIDs that co-occur on both sides of that unit in the parallel corpus and the number of MIDs that occur only on one of its sides, N or M. To express the overall consistency, the mean over all alignment units (C_{AU_i}) of the corpus is calculated.

Table 5.7 depicts the over-all consistency values (columns two and five) starting from lexicon $\mathcal{L}(0)$ (only validated cognates) up to lexicon $\mathcal{L}(4)$ for all target languages (improvements after that step are only marginal, cf. Table 5.6). When processing the English-Spanish corpus, consistency is already about 38%, only considering cognates using the C measure. This surprisingly high value is due to the high amount of overlapping medical terms in different Western European languages. Adding those entries acquired from bootstrapping the same corpus, consistency climbs to a maximum of 46%. As a reference item, the processing of an English-German corpus, which is also derived from UMLS, yields 58% consistency (keeping in mind that English and German lexicons were generated manually and provide a real good coverage (cf. evaluation results in Section 8.2). For English-French, consistency ranges from 40% (only cognates) and 52% (after four bootstrapping cycles) to 57% when including additional manual entries, whilst for English-Swedish, 58% consistency is reached after the automatic acquisition and 63% after the manual insertion of 2,811 additional subwords.

The processing of Spanish-French, Spanish-Swedish, and French-Swedish is particularly interesting, since the underlying corpora were not involved at all in the lexical acquisition process. With consistency starting from 30% for cognates (Spanish-French), 42% is reached after four cycles of generating the non-English lexicons by processing parallel corpora aligned to English only.

Lexicon	C	Cov.(%)	Ident.(%)	C	Cov.(%)	Ident.(%)
	English-Spanish (n = 60,526)			Spanish-French (n = 13,158)		
$\mathcal{L}(0)$	37.7	85.5	7.9	30.4	61.9	15.1
$\mathcal{L}(1)$	45.7	90.8	11.9	40.8	74.6	22.8
$\mathcal{L}(2)$	46.1	91.0	12.2	41.9	75.2	23.6
$\mathcal{L}(3)$	46.1	91.0	12.2	42.0	75.3	23.8
$\mathcal{L}(4)$	46.1	91.0	12.2	42.0	75.3	23.8
$\mathcal{L}_{ALL}$	**53.0**	**92.2**	**13.5**	**47.2**	**78.6**	**26.8**
	English-French (n = 17,130)			Spanish-Swedish (n = 8,993)		
$\mathcal{L}(0)$	39.9	73.6	16.6	23.8	51.6	9.8
$\mathcal{L}(1)$	50.9	84.0	24.8	42.4	73.2	23.2
$\mathcal{L}(2)$	52.0	84.3	25.2	43.5	74.5	23.5
$\mathcal{L}(3)$	52.1	84.4	25.4	43.9	74.6	24.0
$\mathcal{L}(4)$	52.1	84.4	25.4	44.1	74.6	24.2
$\mathcal{L}_{ALL}$	**57.4**	**86.8**	**27.2**	**49.8**	**78.7**	**26.0**
	English-Swedish (n = 10,953)			French-Swedish (n = 6,713)		
$\mathcal{L}(0)$	30.8	56.7	16.3	25.8	51.3	11.5
$\mathcal{L}(1)$	55.5	80.4	38.2	42.9	71.8	23.9
$\mathcal{L}(2)$	57.6	81.5	40.4	44.6	72.9	24.6
$\mathcal{L}(3)$	57.7	81.5	40.4	44.6	73.0	24.6
$\mathcal{L}(4)$	57.8	81.6	40.8	44.9	73.2	24.8
$\mathcal{L}_{ALL}$	**63.1**	**86.3**	**45.3**	**50.0**	**78.5**	**26.7**

Table 5.7: Indexing Consistency (C), Coverage (Cov.) of Lexicons and Number of Identical Indexes (Ident.) at each Stage of Lexicon Generation.

For Spanish-Swedish (French-Swedish, respectively), after four bootstrapping cycles, consistency reaches 44% (45%), with an additional boost of five percentage points after enhancing the Spanish, Swedish and French lexicons by manual generated entries.

Lexical coverage was measured by counting those cases in which at least one MID occurs on both sides of the alignment units considered. This is particularly interesting from the cross-language information retrieval perspective. For Spanish cognates only ($\mathcal{L}(0)$ in Table 5.7), (incomplete) alignments to English can be observed for 86% of the corpus. This value increases to 91% after four runs of bootstrapping the Spanish lexicon, and for English-French, coverage reaches 84% (for English-Swedish 82%). After manually enhancing the lexicons with additional entries ($\mathcal{L}_{ALL}$), coverage increases to 86% for English-Swedish, up to 92% for English-Spanish. For Spanish-French, Spanish-Swedish, and French-Swedish, surprisingly enough, coverage yields 73% to 75% after the automatic acquisition, and 79% after manual lexical enhancements. Again, as a reference value, the processing of the English-German corpus yields 91% coverage.

The number of cases in which both sides are indexed identically, are depicted in Table 5.7, Columns four and seven. The reference data for these values is 30% for English-German. Remarkably, the number of identical indexes is very high for English-Swedish (45%), when compared to the other language pairs. This can be explained by the fact that the relatively imprecise data coming from ICD and ICPC is missing for Swedish in the UMLS Metathesaurus.

5.5 Discussion

The rise of the empirical paradigm in the field of machine translation is, to a large degree, due to the wide-spread availability of parallel corpora (Brown et al., 1990). They also constitute an important resource for the automated acquisition of translational lexicons (Turcato, 1998). Most approaches to multilingual lexical acquisition employ statistical methods, such as context vector comparison (Rapp, 1999; Widdows et al., 2002; Déjean et al., 2002) or mutual information (Fung, 1998) and

require a seed lexicon of trusted translations. Koehn & Knight (2002) derived such a seed lexicon from German-English cognates which were selected by using string similarity criteria (a method also favored by Ribeiro et al. (2001)). Barker & Sutcliffe (2000) propose an alternative *generative* approach where Polish cognate candidates are created from an English word list using string mapping rules, an approach to cognate mapping also discussed by MacWhinney (1995) for 2nd language acquisition of human learners.

The second issue concerns the processing of suitable corpora. Whilst Widdows et al. (2002) deal with parallel German-English corpora to enrich an existing multilingual lexicon (also taken from the UMLS Metathesaurus), Fung (1998), Rapp (1999) and Déjean et al. (2002) propose methods that require only weaker comparable corpora (cf. (Fung, 1998) for a linguistic distinction between both types of corpora). Furthermore, Déjean et al. (2002) incorporate hierarchical information from an external thesaurus (MeSH, 2005) for combining different evidence for lexical acquisition. Cheng et al. (2004) as well as Zhang & Vines (2004) propose co-occurrence-based methods to automatically extract word translations from mixed-language texts which are dynamically made available trough common Web search engines.

Here, in contradistinction to these precursors, a fully heuristic method for acquiring translations of subwords is proposed instead of using statistics. This is made possible by the availability of relatively large and well aligned parallel corpora, as provided within the UMLS Metathesaurus.

Chapter 6

Cross-Lingual Resolution of Acronyms

The understanding of abbreviations in biomedical texts is very important for natural language processing applications, such as information extraction (Friedman & Hripcsak, 1999) or information retrieval systems (Hersh, 2002). This is witnessed, in particular, for protein and gene expressions from biomedical texts (Fukuda et al., 1998), as well as the relations between them (Blaschke et al., 1999). Those expressions frequently consist of acronyms, but their definitions in the text might differ from the ones found, e.g., in an external database, such as ARGH, AcroMed, or SaRAD (Wren & Garner, 2002), cf. Wren et al. (2005) for an overview.

Multiple expansions for a unique acronym, or multiple acronyms for a unique term, will lead to difficulties when trying to match natural language expressions to a standardized vocabulary such as the UMLS or MeSH (Zeng & Cimino, 1996; Aronson et al., 2000; Aronson, 2001; Zweigenbaum et al., 2001; Markó et al., 2003; Markó et al., 2004a; 2006c). In an information retrieval scenario, unresolved acronyms will possibly lead to a loss of precision: Does *"AD"* refer to *"Alzheimer's Disease"* or to *"allergic dermatitis"*? The problem of ambiguity becomes even harder, when multilingual documents are made available to a search interface, which is the case for most Web search engines. In this case, the acronym *"AD"* may have the German expansion *"atopische Dermatitis"*, Spanish *"aurícula derecha"*,

Portuguese *"água destilada"*, and many more. On the other Hand, the German acronym equivalent to *"Alzheimer's Disease"* is *"AK"* (*"Alzheimer Krankheit"*) or *"MA"* (*"Morbus Alzheimer"*) and for Spanish *"EA"* (*"enfermedad de Alzheimer"*).

There has been extensive research on the automatic extraction of short-form/long-form pairs (abbreviations and acronyms mapped to their expansions/definitions) within one language (Adar, 2004; Chang et al., 2002; Pustejovsky et al., 2001; Schwartz & Hearst, 2003; Wren & Garner, 2002). Different ways how abbreviations and acronyms are actually used in written (medical) language have been studied (Liu et al., 2002). However, little attention has been paid on how acronyms and their associated long-forms behave across languages (Hahn et al., 2005a; Markó et al., 2005b; 2006e).

6.1 Algorithm for Acronym Extraction

Schwartz & Hearst (2003) offer a simple and fast algorithm for the extraction of abbreviations and their definitions. The algorithm achieves 96% precision and 82% recall on a standardized test collection, thus, performs at least as good as other existing approaches (Adar, 2004; Chang et al., 2002; Pustejovsky et al., 2001; Wren & Garner, 2002). The source code (in Java) is made available on the Web.[1]

Generally, the process of identifying abbreviations and their full forms can be seen as a two-step procedure: the extraction of possible short-form/long-form (SF-LF) pairs and the validation of SF-LF terms among the list of possible candidates in a sentence.

6.1.1 Extraction of possible SF-LF terms

SF-LF pairs are identified by the adjacency to parentheses. The two basic patterns *LF (SF)* and *SF (LF)* are thereby distinguished. A short form has the following characteristics: it contains between 2 and 10 characters, it has a maximum of two words, at least one character is a letter and its first character is alphanumeric. The

[1] `http://biotext.berkeley.edu/software.html`

long form must immediately appear before or after the corresponding short form and the maximum number of words is constrained by

$$min(|A| + 5, |A| * 2)$$

with |A| being the number of characters in the corresponding SF (a heuristics originally proposed by Park & Byrd (2001) that is also used in recent work by Kokkinakis & Dannélls (2006)). In practice, the first pattern *LF (SF)* proved to occur more frequently. Only if a criterion for an *LF (SF)* pattern is not fulfilled (e.g., more than two words inside the parentheses), the second pattern *SF (LF)* is tried.

6.1.2 Identifying the correct SF-LF term

A set of simple rules is used to identify the correct SF-LF pair out of a set of possible candidates. Most importantly, each character in the short form must match a character in the long form. Characters of the short form must appear in the same linear order as in the long form. Furthermore, the first character of the SF has to be the same in the LF. Finally, all LFs are removed which are shorter than the corresponding SF, or which include the corresponding SF within one of their single words.

6.2 Extracting Biomedical Acronyms

In order to acquire acronyms together with their definitions from biomedical texts heterogeneous Web sources were taken, including MEDLINE abstracts (cf. Table 5.3 in Section 5.1.1.1). With over 250 million words the derived English corpus was much larger than those for the other languages involved (37 millions for German, 14 millions for Portuguese, and 11 millions for Spanish).

Using the algorithm described above, over 1.2m abbreviations were collected for English, together with their long forms (cf. Table 6.1). 31,750 pairs were retrieved for German, 8,029 for Portuguese, 7,675 for Spanish, 3,886 for French, and 266 for Swedish. In contradistinction to the other languages, the English corpus included a large number of expert-level MEDLINE abstracts. As a consequence, every 200th

Language	Corpus Tokens	Acronyms	
English	250,240,401	1,253,318	(0.5%)
MSI-Covered		1,108,921	(88.5%)
German	37,715,960	31,750	(0.08%)
MSI-Covered		29,477	(92.8%)
Portuguese	13,904,790	8,029	(0.06%)
MSI-Covered		7,070	(88.1%)
Spanish	11,460,598	7,675	(0.07%)
MSI-Covered		4,051	(52,8%)
French	4,166,108	3,886	(0.09%)
MSI-Covered		2,603	(67,0%)
Swedish	2,480,573	266	(0.01%)
MSI-Covered		177	(66,5%)

Table 6.1: Corpus and Acronym Extraction Statistics

token in the collection was classified as an acronym. For the other languages (for which the corpora included a higher amount of consumer information), this ratio is much smaller (0.01 to 0.09 percent of the corpora), in particular for Swedish, for which the corpus did not contain any MEDLINE-related abstracts.

After the acquisition of SF-LF pairs, the long forms were normalized to lower-case characters, whilst case sensitivity was kept for short forms, in contrast to previous work (Hahn et al., 2005a; Markó et al., 2005b). The reason for this is that in biology, protein and gene names are differentiated by defined upper- and lower-case characters, and subtle discriminations of referenced species are based on the different use of case. Furthermore, the character normalization of short forms such as *"MG"* and *"mg"* would cause unnecessary ambiguity when resolving to, e.g., *"myasthenia gravis"* or *"milligram"*.

The long forms were then processed with the morpho-semantic indexing (MSI) procedure as described in Section 3.2. Upon prior manual inspection of document samples, it has been observed that English long forms also frequently occur in

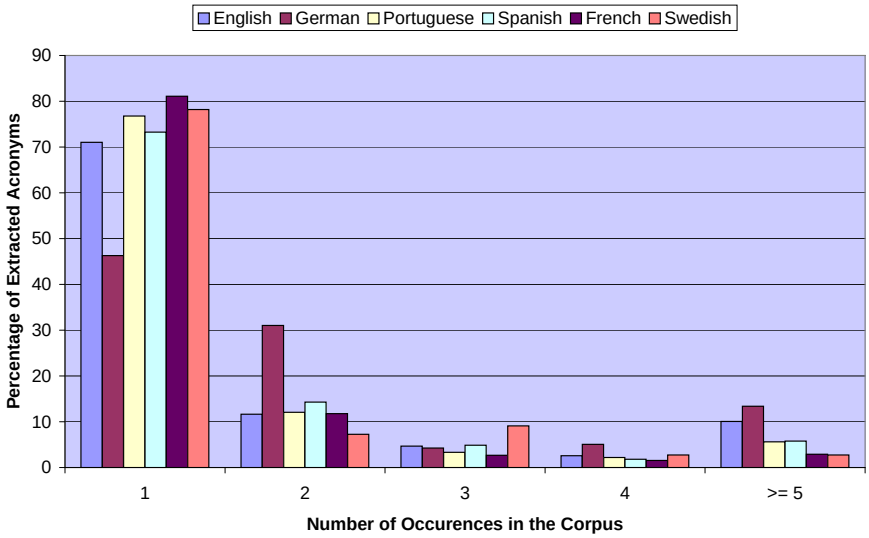

Figure 6.1: Distribution of SF-LF Occurrences in each Corpus

German, Portuguese, and Spanish texts. Therefore, a decision had to be made which lexicon to use for MSI. Therefore, the long forms were segmented using every language-specific subword lexicon involved. Afterwards, those language hypotheses were kept for which the underlying lexicon yielded complete lexical coverage with regard to the specific long form. If there were more than one remaining language hypothesis, the document language (if not English) was preferred over English.

This procedure led to over one million SF-LF form pairs covered by the MSI procedure for English (89%), 29,477 (93%) for German, 7,070 (88%) for Portuguese, 4,051 (53%) for Spanish , 2,603 (67%) for French and 177 (67%) for Swedish (cf. Table 6.1). In the following, this (sub-) set of extracted abbreviations is focused on only.

Figure 6.1 gives an impression of how frequent distinct SF-LF pairs occurred in the corpora considered, for each language condition. 46% to 81% of all acronyms extracted occurred only once, 7% to 31% appeared twice, whilst five or more occurrences were found for 3% to 13% of all SF-LF pairs.

As depicted in Table 6.2 (Column 2), 235,076 unique SF-LF pairs were generated for English, 4,732 for German, 3,983 for Portuguese, 1,993 for Spanish, 1,793 for French, and 110 for Swedish. Column 3 of the table shows the average number of

Language	Surface		MSI	
	Unique	Ratio	Unique	Ratio
English	235,076	4.72	214,590	5.17
German	4,732	6.23	3,970	7.43
Portuguese	3,983	1.78	3,674	1.92
Spanish	1,993	2.03	1,880	2.16
French	1,793	1.45	1,727	1.51
Swedish	110	1.61	98	1.81

Table 6.2: Effects of Morpho-semantic Normalization in Terms of Unique SF-LF Pairs and Tokens per Type

corpus occurrence for each unique SF-LP pair. After the morpho-semantic normalization of long forms, the number of unique SF-LF pairs decreased as expected, e.g. to 214,590 for English. Accordingly, the number of tokens per type increased, as depicted in the fifth column of Table 6.2. As an example, morpho-syntactic variants in long forms such as in *"CTC"- "computed tomographic colonography"* and *"CTC"- "computed tomography colonography"* were unified. Hence, additional evidence for the validity of such an extracted SF-LF pair increases.

6.3 Results

Extracted acronym-definition pairs were examined under two conditions. Firstly, they were analyzed regarding their behavior within one particular language. Afterwards, relying on the language-independent representation of long forms using MorphoSaurus, cross-lingual constellations of SF/LF pairs were analyzed in a more detailed way.

6.3.1 Intra-Lingual Phenomena

Two basic phenomena have to be considered when inspecting the results for one given language. At first, one short form can have multiple long forms, and, secondly, one

Language	**SF**	**Average**	
		Surface	MSI
English	90,518	2.60	2.37
German	3,206	1.48	1.24
Portuguese	2,540	1.57	1.45
Spanish	1,394	1.43	1.35
French	1,362	1.32	1.27
Swedish	87	1.26	1.13

Table 6.3: Number of Long Forms for Each Short Form (SF)

Language	**Surface**		**MSI**	
	LF	Average	LF	Average
English	199,633	1.18	170,185	1.26
German	4,598	1.03	3,772	1.05
Portuguese	3,845	1.04	3,414	1.08
Spanish	1,941	1.03	1,793	1.05
French	1,755	1.02	1,665	1.04
Swedish	109	1.01	97	1.01

Table 6.4: Number of Short Forms for each Long Form (LF)

long from can have multiple short forms. An example for a SF ambiguity is given with *"ABM"* mapped to *"acute bacterial meningitis"* or to *"adult bone marrow"*. Table 6.3 shows the numbers of different long forms for each short form, both for the baseline condition (surface forms) and the MSI condition. For English, 90,518 unique short forms were extracted. The average number of long forms associated to unique SFs decreases from 2.60 to 2.37 for MSI, as expected. The same relationship can also be observed for the other languages considered.

The second phenomenon is also observable in all languages involved in the experiments. For example, the noun phrase *"acid phosphatase"* has nine different ab-

breviations in the English corpus processed: *"AcP"*, *"acPAse"* *"ACP-ase"*, *"Acph"*, *"ACPT"*, *"AP"*, *"APase"*, *"AphA"*, and *"APs"*. Table 6.4 depicts the numbers describing this phenomenon. For English, a total of 199,633 different long forms were extracted, embedded in 235,076 different SF-LF pairs (cf. Table 6.2). Thus, each LF is associated to 1.18 SFs, on the average. For the MSI condition, there are less different long forms, hence, the ratio slightly increases, for all languages.

6.3.2 Inter-Lingual Phenomena

6.3.2.1 Identical SF-LF Pairs

The first observation is that quite often SF-LF pairs are identical across languages on the surface level. Especially common or technical English terms also appear in other languages, such as *"WHO"* and its expansion *"World Health Organization"*, *"PCR"* and its definition *"polymerase chain reaction"*, or *"IL"* associated to *"interleukin"*. In numbers (cf. Table 6.5, Column 2), up to 163 identical SF-LF pairs for Portuguese-Spanish, 189 for English-French, and 478 for English-German have been found, while language pairs not related to English also may contain some English SF-LF pairs. Consequently, foreign-language SF-LF pairs should also be included in a language-specific lexicon for properly applying lexicon-based NLP-tools.

6.3.2.2 Identical SF, Different LF

One way of identifying possible translations of long forms is to collect those long forms, which are connected to a unique short form at the surface level. For example, if an English document contains *"WHO"*- *"World Health Organization"* and a German document contains *"WHO"*- *"Weltgesundheitsorganisation"*, the long forms can be regarded as possible translations of each other. For English-Portuguese, 129,957 of these pairs can be extracted and for English-German, there are 78,761 of these hypothesized translations (Table 6.5, Column 3). Of course, these sets also contain syntactic variants and a large number of false positives, since short forms are used differently across languages. Therefore, the focus is switched to the interlingual layer of long form representations.

	Surface		MSI	
Language Pair	**I(SF) I(LF)**	**I(SF) D(LF)**	**I(SF) T(LF)**	**D(SF) T(LF)**
English-German	478	78,761	1,016	2,540
English-Portuguese	154	129,957	371	3,665
English-Spanish	165	82,565	309	2,044
English-French	189	54,833	312	1,490
English-Swedish	28	2,978	57	153
German-Portuguese	33	2,219	67	268
German-Spanish	28	1,452	62	152
German-French	27	1,081	69	95
German-Swedish	11	202	26	15
Portuguese-Spanish	163	3,203	255	174
Portuguese-French	15	2,041	93	89
Portuguese-Swedish	0	75	1	18
Spanish-French	7	1,131	55	41
Spanish-Swedish	1	41	4	13
French-Swedish	2	50	7	7
Total	1,301	360,589	2,704	10,764

Table 6.5: Statistics on Cross-Lingual Acronym Extraction: Results for Identical (I), Different (D) and Translations (T) of Short Forms (SF) and Long Forms (LF)

6.3.2.3 Identical SF, Translation of LF

In this condition, those cases were examined, in which short forms were identical and long forms were different at the surface level, but identical at the interlingual layer, comparing SF-LF pairs extracted from the different source corpora. As a result, lists of bilingually aligned terms were acquired, such as English *"acute lymphatic leukemia"* linked to the German *"akute lymphatische Leukämie"* via the shared short term *"ALL"*. As an example, 1,016 translations were generated for English-German using this heuristics (cf. Table 6.5, Column 4).

6.3.2.4 Different SF, Translation of LF

In this scenario, those cases were analyzed, for which the long forms were identical or translations of each other (identical at the interlingua layer), but with different short forms. This captures interesting constellations such as English *"AIDS"* (*"acquired immune deficiency syndrome"*) aligned to Portuguese or Spanish *"SIDA"* (*"síndrome de inmunodeficiencia adquirida"*). In total, up to 3,665 of these translations were collected for English-Portuguese (Table 6.5, Column 5).

6.4 Lexicon Integration

In order to enrich the existing lexicons with acronyms automatically, the quality of the derived associations of short forms to long forms had to be ensured. With 96% precision, as measured by Schwartz & Hearst (2003) on a standardized test set, over 9,000 false positives can be expected in the set of unique SF-LF pairs, only considering English (cf. Table 6.2). Furthermore, since MorphoSaurus focuses on Cross-Language Information Retrieval (Markó et al., 2005c; 2005f) and multilingual text classification (Markó et al., 2003; Hahn et al., 2004a), cross-lingual mappings of lexical entries are of particular interest. Both challenges are met by a simple heuristics, based upon the idea that *two languages are more informative than one* (Dagan et al., 1991). Hence, those extracted SF-LF pairs were incorporated in the available subword lexicons, for which the long form is a translation of, at least one, another long form in a different language (mapped on the interlingua layer).

Language	Initial Size	New Acronyms
English	22,561	62,236
German	23,976	2,932
Portuguese	14,984	2,195
Spanish	10,936	1,275
French	7,812	834
Swedish	10,281	80
Sum	90,550	69,552
Total	160,102	

Table 6.6: Subword Lexicon Size

Thus, those pairs were collected for which the number of occurrences are depicted in Column 4 and 5 in Table 6.5.

As a result, an intersection of 3,024 English SF-LF forms were obtained, 1,281 for German, 1,342 for Portuguese, 774 for Spanish, 575 for French, and 67 for Swedish (a total of 7,063). For the monolingual mapping of short forms to long forms, those language-specific SF-LF pairs were collected, which occur at least twice on the layer of the interlingua (cf. Table 6.2, right).

In the end, the lexicon size for the specific languages increased from initially 90,550 entries to 160,102 lexical items (cf. Table 6.6).

Formally, the lexicon integration was realized by adding the acronym to the subword lexicon as an invariant and by creating (a) unique MID(s) for each of the associated long forms in the thesaurus, to which the new lexeme was linked. For example, different readings for the new subword entry *"AD"* were firstly encoded by using the distinct MIDs $\#AD1$ and $\#AD2$:

$$l_{46} = (AD, IV, \#AD1, EN, d_1)$$
$$l_{47} = (AD, IV, \#AD2, EN, d_1)$$

Afterwards, the relations

$$\{(\#AD1, \#atop), (\#AD1, \#dermat), (\#AD1, \#inflamm)\} \in expandsTo$$

and

$$\{(\#AD2, \#alzheimer), (\#AD2, \#diseas)\} \in expandsTo$$

were added to the thesaurus, covering the different interpretations *"atopic dermatitis"* and *"Alzheimer's Disease"*, respectively. Then, for correctly identifying acronyms during the morpho-semantic processing of input texts, the lexicon look-up is performed in a preprocessing step, before transforming word characters to lower-case (cf. Section 3.2.3.1).

6.5 Discussion

Several different techniques for the automatic extraction of acronyms and their definitions from biomedical text (particularly from MEDLINE abstracts) have been developed up until now (Pustejovsky et al., 2001; Chang et al., 2002; Wren & Garner, 2002; Schwartz & Hearst, 2003; Adar, 2004). Comprehensive databases with millions of entries are provided by different research groups (Pustejovsky et al., 2001; Wren & Garner, 2002; Chang et al., 2002; Adar, 2004). They adopt similar sorts of heuristics such as identifying and processing parenthetical phrases within texts. Some of them use stemming (Pustejovsky et al., 2001; Adar, 2004), and/or apply term normalization routines to their abbreviations and full forms (Pustejovsky et al., 2001; Chang et al., 2002; Adar, 2004; Okazaki & Ananiadou, 2006). Pustejovsky et al. (2001) incorporate a shallow parsing approach. A general overview of the four large databases and their algorithms can be found in the work of Wren et al. (2005).

The approach for the multilingual alignment of acronyms and their definitions as described in this chapter is tied up to the research from these precursors. By translating extracted long forms into an interlingual layer, an approach which has not been exploited so far, acronyms and their definitions are made comparable across different languages with a high coverage.

Chapter 7

Subword Sense Disambiguation

Automatic word sense disambiguation (WSD) is one of the most challenging tasks in natural language processing, and, therefore, has been a long-term concern for computational linguistics (cf. Ide & Véronis (1998) and Kilgarriff & Palmer (2000)). Since the mapping from lexical forms to senses is usually 1:n, multiple semantic readings for a word have to be considered and, at best, reduced to a single one on a routine basis. Typically, the source for such multiple meaning assignments are lexical databases, dictionaries or thesauri, the most prominent example being WordNet (Fellbaum, 1998). WSD approaches can then broadly be distinguished into symbolic ones (Voorhees, 1993) and corpus-based ones (Gale et al., 1993). Although the latter became popular due to the increasing availability of large machine-readable corpora, Dagan & Itai (1994) point out that corpus-based WSD requires manually sense-tagged training data (supervised WSD). Brown et al.'s (1991) usage of bilingual corpora is certainly a good idea to avoid manual tagging of training material but such corpora are only available for a limited number of domains. Dagan & Itai (1994) then came up with the idea that WSD for machine translation might complement bilingual dictionaries with monolingual corpora, which are much easier to obtain.

The following methodology tries to combine the best of both worlds. On the one hand, it adheres to an unsupervised approach to WSD because it requires no human intervention. On the other hand, it takes advantage from already existing lexical and textual resources in terms of multilingual thesauri, as well as unaligned, though

comparable corpora for six different languages, *viz.* English, German, Portuguese, Spanish, French, and Swedish (for a linguistically motivated distinction of parallel and comparable corpora, cf. Fung (1998)).

The proposed approach rests on the idea that although multiple senses can be attributed to the *same* single lexical item in one language, these senses usually are denoted by *different* lexical items in other languages (Ide, 2000). As an example, consider the German lexical form *"Krebs"*, which can either refer to *"cancer"* or *"crab"*. Given comparable (i.e., topically related) corpora, the context they provide helps in deciding which variant is more likely to be intended. At the level of the same language, it may also be helpful to consider non-ambiguous synonyms, hypernyms or hyponyms such as the German word *"Karzinom"* (*"carcinoma"*). Context words of the latter type can then be used for identifying the proper sense of the given polysemous item. But multilingual disambiguation may not always be so straightforward. Consider, e.g., the English lexical item *"patient"*, which has (at least) two different meanings. As a noun it refers to a human, as an adjective it has a completely different meaning. Unfortunately, there is no (unambiguous) synonym to the first reading. Even the translation to French, *"patient"*, is also ambiguous and covers the same meaning facets. However, the German translation, (*"Patient"*), has only one meaning, *viz.* a human in need of medical treatment (the German translation of the adjective *"patient"* yields *"geduldig"*).

In the following, it will be shown that this interrelation, i.e. different senses of a given word tend to have different translations in other languages, can be used for collecting better evidence for automatic, unsupervised word sense disambiguation (Markó et al., 2005e).

7.1 Combining Multilingual Evidence for WSD

For the experiments, the medical corpora introduced in Section 5.1.1.1, Table 5.3 are used once again. The collections were split into training (75%) and test sets

Language	Word Tokens	MID Tokens	Ambiguous MIDs	Number of Readings
English	187,992,247	145,175,273 (77.2%)	17,281,993 (11.9%)	85,913,094 (avg. 5.0)
German	29,046,282	16,125,018 (72.7%)	2,056,470 (12.8%)	4,721,794 (avg. 2.3)
Portuguese	9,864,434	7,336,285 (74.4%)	732,421 (10.0%)	1,683,744 (avg. 2.3)
Spanish	10,758,234	7,384,183 (68.6%)	347,571 (4.7%)	804,006 (avg. 2.3)
French	3,116,236	2,374,537 (76.2%)	152,555 (6.4%)	375,522 (avg. 2.5)
Swedish	2,300,565	1,099,063 (47.8%)	179,370 (16.3%)	411,757 (avg. 2.3)
Mixed	40,788,650	30,568,341 (74.9%)	3,549,487 (11.6%)	15,896,109 (avg. 4.5)

Table 7.1: Training Corpus Statistics

(25%)[1], resulting in 2.3 million training tokens for Swedish, up to 188 million tokens for English (cf. Table 7.1, second column). So, the sizes of the training corpora vary significantly across the languages considered due to their unequal availability. For Portuguese, Spanish, French and Swedish the amount of training data is relatively small compared to other work on data-driven WSD (e.g., the 25 million words corpus used by Dagan & Itai (1994) or the 50 million words corpus used by Schütze (1992)).

[1]Since the context of words have to be preserved, documents of the collections (not phrases or words) were split. As a consequence, the relation between the number of tokens of training and test sets is not exactly 75% vs. 25%. In fact, for Swedish and Spanish, the ratio differs considerably because of highly varying document sizes.

7.1.1 Training the Classifier

Using the subword lexicons extended with acronym definitions as described in the last chapter (cf. Table 6.6), the training corpora were processed by MorphoSaurus, resulting in the interlingual content representation of original texts. For the experiments described in the following, lexical remainders due to incomplete segmentations of original words are ignored (cf. Section 3.2.3.2).

Furthermore, in order to test the influence of multilingual sources, a mixed training set was built by taking the sixth part of each of the (morpho-semantically normalized) (six) different language-specific training corpora.

This led to 145 million equivalence class identifiers (MIDs) for English, corresponding to 77% of the original number of tokens (cf. Table 7.1, third column). Similar ratios were observed for German and Portuguese, while for the automatically acquired lexicons for Spanish, French, and Swedish the numbers of resulting MIDs differ significantly. For the mixed training corpus, the ratio averages 75%. The relative number of ambiguities in the resulting representations range from 5% for Spanish up to 16% for Swedish (Table 7.1, fourth column). Except for English, where a substantial amount of the corpus is comprised of MEDLINE abstracts containing (highly ambiguous) acronyms and abbreviations (cf. the previous chapter), the average number of readings for each ambiguity is relatively constant for each training condition (ranging from 2.3 to 2.5, cf. Table 7.1, fifth column[2]).

Finally, evidence for the test phase was collected by counting co-occurrences of equivalent class MIDs within a window of ± 2 *unambiguous* MIDs (a size already proposed in early experiments by Kaplan (1955)). Ambiguous MIDs are completely ignored in the training phase. Resulting counts of co-occurrence patterns are then stored separately for each of the training conditions (English, German, Portuguese, Spanish, French, Swedish and mixed).

[2]For comparison, Dagan & Itai (1994) identified 3.3 "senses" per word defined as possible translations to a target language (both for German-English and Hebrew-English).

Language	Word Tokens	MID Tokens	Ambiguous MIDs	Number of Readings
English	62,248,154	48,349,369 (77.7%)	5,755,653 (11.9%)	28,615,648 (avg. 5.0)
German	8,669,678	5,498,861 (63.4%)	702,713 (12.8%)	1,615,248 (avg. 2.3)
Portuguese	4,040,356	2,996,010 (74.2%)	296,887 (9.9%)	680,675 (avg. 2.3)
Spanish	702,364	477,269 (68.0%)	21,666 (4.5%)	50,793 (avg. 2.3)
French	1,049,872	799,446 (76.2%)	51,335 (6.4%)	126,430 (avg. 2.5)
Swedish	180,008	92,406 (51.3%)	13,180 (14.3%)	29,823 (avg. 2.3)

Table 7.2: Test Corpus Statistics

7.1.2 Testing the Classifier

The test collection comprised 180,000 tokens for Swedish, up to 62 million tokens for English. The data exhibits similar ratios of MIDs after the morpho-semantic processing as seen in the training collections (cf. Table 7.2, second and third column). The number of ambiguous MIDs range from 5% for Spanish up to 14% for Swedish, with the same average number of meanings as in the training collections.

For testing, a well-known probabilistic model was used, the *maximum likelihood estimator* (Manning & Schütze, 1999). For each ambiguous subword at position k with n readings, resulting in a sequence of equivalence class identifiers, $MID_{1,k}$, $MID_{2,k}$, ..., $MID_{n,k}$, examine the window of $\pm w$ surrounding items. Then, with $f(x, y)$ denoting the frequency of co-occurrence of the MIDs x and y in the training corpus, choose that particular MID_i $(1 \leq i \leq n)$ for which the probability

$$P_{MID_i} = \sum_{j=1}^{w} \frac{f(MID_{i,k}, MID_{i,k-j}) + f(MID_{i,k}, MID_{i,k+j})}{\sum_{m=1}^{n} f(MID_{m,k}, MID_{m,k-j}) + f(MID_{m,k}, MID_{m,k+j})}$$

is maximal. If there is no observable maximum following this procedure, disambiguation fails.

Primarily, the coverage of the classifier was measured in this experiment, rather than its accuracy, since, unfortunately, the only available test collection for biomedical WSD (Weeber et al., 2001) is not suitable for the needs, due to different target categories (weak semantic types as encoded in the UMLS, rather than MIDs).[3]

7.1.3 Results

Table 7.3 depicts the test results after the disambiguation of ambiguous subwords using monolingual (column three and four) and multilingual (column five and six) training texts. Just as in the training phase, a window of two surrounding items is examined (rows four to nine). Another typical context span for WSD described in the literature is a window of six items (cf. Ide & Véronis (1998)). Coverage values for this condition are shown in rows 11 to 16.

Considering $w = \pm 2$ (i.e. two tokens surrounding an ambiguous MID on each side) in the monolingual training scenario, the ratio of ambiguous MIDs declines to 3.0% for Swedish, down to 0.4% for English. Hence, between 77% and 97% of all ambiguous MIDs can be resolved for Swedish (small training set) and English (large training set), respectively. The only source for discriminating word senses in this test condition are synonyms covered by the MorphoSaurus lexicons.

Given this (monolingual) baseline, it has been tested which improvements (if any) can be observed using the same test set and scenario, but incorporating multilingual material in the training. As shown in Table 7.3 (column five and six), for English, only 0.2% of the produced MIDs remain ambiguous, which means that 99% of all ambiguities can be resolved. For German, the benefit comes to a 9.2 percentage points gain, whilst for Swedish the proportion of resolved ambiguities increases

[3]For general language use, the *Brown Corpus* and the *Wall Street Journal* provide taggings with WordNet senses (Ng & Lee, 1996), while SENSEVAL in the first competition round started with Hector senses (Kilgarriff & Palmer, 2000) and only in the second one turned to WordNet senses, as well.

		Monolingual Training		Multilingual Training	
Language	**MIDs**	**Ambiguous**	**Resolved**	**Ambiguous**	**Resolved**
$w = \pm 2$					
English	48,349,369	192,488 (0.40%)	5,563,165 (**96.7%**)	82,686 (0.17%)	5,672,967 (**98.6%**)
German	5,498,861	109,728 (2.00%)	592,985 (**84.4%**)	45,207 (0.82%)	657,506 (**93.6%**)
Portuguese	2,996,010	12,691 (0.42%)	284,196 (**95.7%**)	784 (0.03%)	296,103 (**99.7%**)
Spanish	477,269	3,641 (0.76%)	18,025 (**83.2%**)	121 (0.03%)	21,545 (**99.4%**)
French	799,446	10,160 (1.27%)	41,175 (**80.2%**)	589 (0.07%)	50,746 (**98.9%**)
Swedish	92,406	3,093 (3.35%)	10,087 (**76.5%**)	8 (0.01%)	13,172 (**99.9%**)
$w = \pm 6$					
English	48,349,369	164,924 (0.34%)	5,590,729 (**97.1%**)	43,245 (0.09%)	5,712,408 (**99.2%**)
German	5,498,861	98,264 (1.79%)	604,449 (**86.0%**)	43,543 (0.79%)	659,170 (**93.8%**)
Portuguese	2,996,010	7,565 (0.25%)	289,322 (**97.5%**)	362 (0.01%)	296,525 (**99.9%**)
Spanish	477,269	3,487 (0.73%)	18,179 (**83.9%**)	85 (0.02%)	21,581 (**99.6%**)
French	799,446	9,253 (1.16%)	42,082 (**82.0%**)	421 (0.05%)	50,914 (**99.2%**)
Swedish	92,406	2,899 (3.14%)	10,281 (**78.0%**)	2 (0.00%)	13,178 (**100%**)

Table 7.3: Coverage Statistics after Disambiguation Based on Monolingual and Multilingual Evidence at Different Window Sizes

from 76.5% for monolingual training to 99.9% for multilingual training. Keeping in mind that the size of the mixed training set (41 million tokens) was significantly smaller than the size of the English training collection (188 million tokens), these results are really promising. Another advantage of combining multilingual evidence becomes clear when observing the French and Swedish test scenario. Due to limited availability, the monolingual training corpus were quite small (3.1 and 2.3 million tokens). Including further available training data from other languages than French and Swedish, evidence for disambiguation also transfers from these languages.

Using a span of ± 6 surrounding tokens, it is likely that coverage improves since more evidence is collected, but this benefit comes at the cost of performance (a factor of 3 compared to $w = \pm 2$). In this scenario, even up to 100% of all ambiguous subwords can be resolved (Swedish), with a gain of up to 22 percentage points for the multilingual training condition.

In previous work (Markó et al., 2005e), the accuracy of the proposed approach was examined in detail for English, German and Portuguese. The correct readings of the subwords in question were determined manually, for a random sample of 100 ambiguous cases for each language and test scenario. Since this is a highly difficult and time-consuming task, the random samples usually drawn for these kind of studies are very small. Dagan & Itai (1994) considered 103 ambiguous Hebrew and 54 German words in their study, whereas Schütze (1992) examined only 10 words and Yarowsky (1992) 12 words. Voorhees (1993) circumvents this dilemma by performing an evaluation *in vivo*, i.e., disambiguation results are considered in terms of the overall performance of a particular application, such as information retrieval or machine translation. Such kind of evaluation for subword disambiguation is presented in the next chapter in a Cross-Language Information Retrieval setting.

Using much smaller training collections but the same test scenario in the previous work, the average accuracy amounts to 60% for the monolingual training condition, and 72% for the multilingual condition. These results are in line with current research on WSD (Kilgarriff & Palmer, 2000; Ciaramita et al., 2003).

7.2 Discussion

For automatic word sense disambiguation (WSD), two major sources of information can be identified. Firstly, external knowledge sources, e.g., *symbolic* syntactic, lexical or encyclopedic knowledge organized in machine-readable dictionaries, thesauri or even more sophisticated ontologies are used. Disambiguation can then be achieved by, e.g., computing the "semantic distances" of the target word and context words, i.e., finding chains of connections between words (Ciaramita et al., 2003), or by identifying overlapping edges in IS-A hierarchies, as proposed by Voorhees (1993), both using "world knowledge" encoded in WordNet (Fellbaum, 1998). Romacker et al. (1999) and Romacker & Hahn (2001) describe an integrated approach for resolving different types of ambiguity occurring in natural language processing by relying on explicit lexical, syntactic and semantic knowledge which is made available through an even more expressive (though domain-limited) description logics based system underlying the (med-)SynDiKATe text understanding system (Hahn et al., 2000; 2002a; 2002b).

Secondly, with the availability of large corpora *data-driven* or *corpus-based* WSD methods gained increasing attention (Gale et al., 1993). Encouraging results were achieved with up to 92% precision using unsupervised machine learning methods on a non-standardized testset (Yarowsky, 1992). Brown et al. (1991) introduced a statistical WSD method for machine translation using aligned bilingual corpora as training data. This approach, however, suffers from the limited availability of such corpora, especially for the medical domain on which is the focus here.

To the best of knowledge, Dagan & Itai (1994) were the first to propose a method using co-occurrence statistics (as well as syntactic knowledge) in unaligned monolingual corpora of two languages. Different senses of a word were defined as all its possible translations into a target language (English), using Hebrew-English and German-English bilingual lexicons. They also made use of the observation that *different senses* of a word from the source language are usually mapped to *different words* in other target languages. They report coverage (applicability) of 68% at 91% precision for Hebrew-English and 50% coverage at 78% precision for German-English. Their results were based on sophisticated significance tests

for making disambiguation decisions and then compared to simple *a priori* frequencies. The latter usually serve as a benchmark for comparison with other decision models, such as Bayesian classifiers (Gale et al., 1993; Yarowsky, 1992; Chodorow et al., 2000), mutual information measures (Brown et al., 1991), context vectors (Schütze, 1992), or neural networks (Towell & Voorhees, 1998) (cf. also Leacock et al. (1996) and Lee & Ng (2002) for an overview and Ng & Lee (1996) for an integrated approach). However, taking only *a priori* frequencies into account, precision drops to 63% (Hebrew-English) and 56% (German-English).

The approach described in this chapter differs from these precursors in several ways. First of all, instead of using bilingual dictionaries, multilingual subword lexicons connected to a thesaurus are used and, hence, operate at an interlingua level of semantic representation. Based on a concept-like representation of word meanings, in contrast to language-specific surface forms, associations between those identifiers can be collected across languages, thus getting rid of the need for aligned bilingual corpora. Secondly, the work of Dagan & Itai (1994) focuses on machine translation, thus, also takes syntactic knowledge into account, whilst the MorphoSaurus approach abstracts away from language-specific particularities (and idiosyncrasies). Comparing coverage values from our approach to those proposed by Dagan & Itai (1994) (68%, respectively 50%, see above) the advantages of using an intermediate, interlingual representation become immediately evident. With trainings on monolingual corpora using ±6 surrounding items of the ambiguous subword in focus, coverage using the subword approach already reaches 87% for all languages, in average (cf. Table 7.3). Compiling these corpora to a multilingual training set, applicability increases to an average of 99%.

Limitations of the approach by Dagan & Itai (using bilingual dictionaries) and Brown et al. (1991) (using bilingual corpora) are discussed by Ide & Véronis (1998). The arguments they raise are also relevant to the investigation proposed here: Many ambiguities are preserved in other languages. Whilst the English word *"patient"* has different translations for German, but not for French (see the introduction of this chapter), it is hard to find similar relations for the word *"mouse"*, which has (at least) the same two meanings of *animal* and *device* for German *"Maus"*, Portuguese

"rato", Spanish *"ratón"*, French *"souris"*, Swedish and Danish *"mus"*, Dutch *"muis"* and Polish *"mysz"*. Nevertheless, by way of identifying a language in which there exists such an unambiguous synonym to the many possible polysemous translations this would entirely suffice for collecting cross-language evidence for disambiguating the ambiguous word in any of these source languages.

Chapter 8

Cross-Language Information Retrieval

Medical document retrieval presents a unique combination of challenges for the design and implementation of retrieval engines (cf. Section 2.1 and Section 2.4). The sheer amount of data available in clinical information systems on the one hand or, on the other hand, in the Web (expert and consumer information portals, bibliographic databases, etc.) rules out the reuse of many of the sophisticated retrieval approaches which perform so well under small-scale experimental conditions such as Latent Semantic Indexing (Deerwester et al., 1990) or even more sophisticated probabilistic models (Fuhr, 1992). The reason for this is that no search engine is capable of maintaining high-dimensional document-term vectors ($n \gg 100,000$) for such an enormous volume of documents and high rate of update frequencies.

Other challenges are given by the multilinguality of medical information available, and the heterogeneous user community. So are clinical findings usually reported in a particular native language spoken in the clinicians country whilst there is a strong bias to English regarding scientific literature (cf. the MEDLINE database).

Cross-Language Information Retrieval (CLIR) is a subfield of information retrieval dealing with retrieving information written in a particular language which is different from the language of the user's query. For example, a user may pose a query in French, but retrieves relevant documents written in English (Grefen-

stette, 1998). Approaches to CLIR can broadly be divided into dictionary-based and corpus-based approaches (Oard & Diekema, 1998). While dictionary-based approaches are both time and cost intensive in performing the language transfer (Levow et al., 2005), they face a number of challenges, including dictionary coverage, morphological variant identification, phrase and proper name recognition, as well as word sense disambiguation (Pirkola et al., 2001).

In this chapter, MorphoSaurus with its underlying subword lexicons is used in CLIR settings for the medical domain. In particular, the performance of the manually built English, German and Portuguese lexicons (cf. Section 3.2) is contrasted to the automatically acquired French, Spanish, and Swedish dictionaries (Section 5.3). The interlingua-based retrieval approach is furthermore compared to an alternative method which relies on the direct translation of non-English (German, Spanish, French, Portuguese and Swedish) user queries to English ones for subsequent processing on large English medical document collections. In addition, the contribution of the acronym resolution module (Chapter 6) and the subword disambiguation module (Chapter 7) to the performance of MorphoSaurus-based CLIR settings is analyzed in detail.

8.1 Experimental Setting

The experiments were run on the Ohsumed corpus (Hersh et al., 1994a), which constitutes one of the standard IR testbeds for the medical domain, and the 2006 corpus of ImageCLEFMed (cf. Clough et al. (2005)).

8.1.1 The Ohsumed corpus

As a subset of the MEDLINE database, Ohsumed contains bibliographic information (author, title, abstract, index terms, etc.) of biomedical articles. Considering the title and abstract field (if available) for each bibliographic unit, the set contains 348,566 documents and 26,705,691 tokens, resulting in an average document length of 76.6 tokens.

The Ohsumed corpus was created specifically for IR studies, and its added value

lies in the fact that 106 authentic user queries are available for which the relevant documents in the corpus had been manually assigned (actually 105, because for one query no relevant documents could be found). OHSUMED, thus, constitutes an unique gold standard for information retrieval experiments in the medical domain. The average number of query terms is 5.2. The following is a query from the set: *"Are there adverse effects on lipids when progesterone is given with estrogen replacement therapy?"*.

8.1.2 The IMAGECLEFMED 2006 corpus

IMAGECLEF is the cross-language image retrieval track which was run as part of the Cross Language Evaluation Forum (CLEF) campaign. IMAGECLEFMED evaluates the retrieval of medical images described by text captions based on queries in different languages. The main goal is to improve the retrieval of medical images from heterogeneous and multilingual document collections containing images as well as textual data.

In IMAGECLEFMED 2006, the multilingual image retrieval task is based on a dataset containing images from different types. Casimage[1] and PEIR (Pathology Education Instructional Resource)[2] contain radiology and pathology images. The MIR collection (Mallinckrodt Institute of Radiology)[3] contains clinical case descriptions related to nuclear medicine and PathoPic[4], finally, is a collection comprised of pathology image descriptions. Considering English annotations only, there are 40,709 image descriptions with a highly variable quality within and between the collections. The number of tokens is 1,130,419, thus, the average document length (27.7) is relatively small compared to OHSUMED. There are 30 queries (topics) for which relevance judgments are available. For English queries, the average number of query terms is 5.8. A typical example is: *"images of a frontal head MRI"*.

[1] `http://www.casimage.com/`

[2] `http://peir.path.uab.edu`

[3] `http://gamma.wustl.edu/home.html`

[4] `http://alf3.urz.unibas.ch/pathopic/e/intro.htm`

8.1.3 Approaches to CLIR

The OHSUMED corpus and the IMAGECLEFMED subset considered here contain only English-language documents. This raises the question of how such collections (or, e.g. MEDLINE) can be accessed from other languages as well. It is a realistic scenario, because, unlike in sciences with English as a *lingua franca*, among medical doctors native languages are dominant in their education and everyday practice and English medical sublanguage capabilities are often quite limited. Otherwise they might resort to translating their native-language search problem to English with the help of current Web technology, e.g., an automatic translation service available in a standard Web search engine. The translation process could additionally be supported by the multilingual UMLS Metathesaurus (UMLS, 2005) which currently supports (with considerable differences in coverage, cf. Section 5.2) German, French, Spanish, Portuguese, Swedish, and many others. Relying on the quality of the translation, this procedure then reduces the cross-language retrieval problem to a monolingual one.

As an alternative, MORPHOSAURUS is used to underpin medical cross-language retrieval. Both approaches will then be evaluated on the same query and document set. As the baseline for the experiments, a standard retrieval system is provided, operating with the Porter stemmer (Porter, 1980) and stop word elimination[5] so that the system runs on (original) English documents with (original) English queries.

In the following experiments, the original English queries were translated into Portuguese, German, Spanish, French and Swedish by medical experts (native speakers of those languages, with a very good mastery of both general and medical English). In Figure 8.1, the result of processing the first query of the OHSUMED collection and an extract of one retrieved document illustrate the two alternative approaches discussed in the following (bold terms co-occur in queries and the document fragment).

[5]The stemmer is available on `http://www.snowball.tartarus.org`.

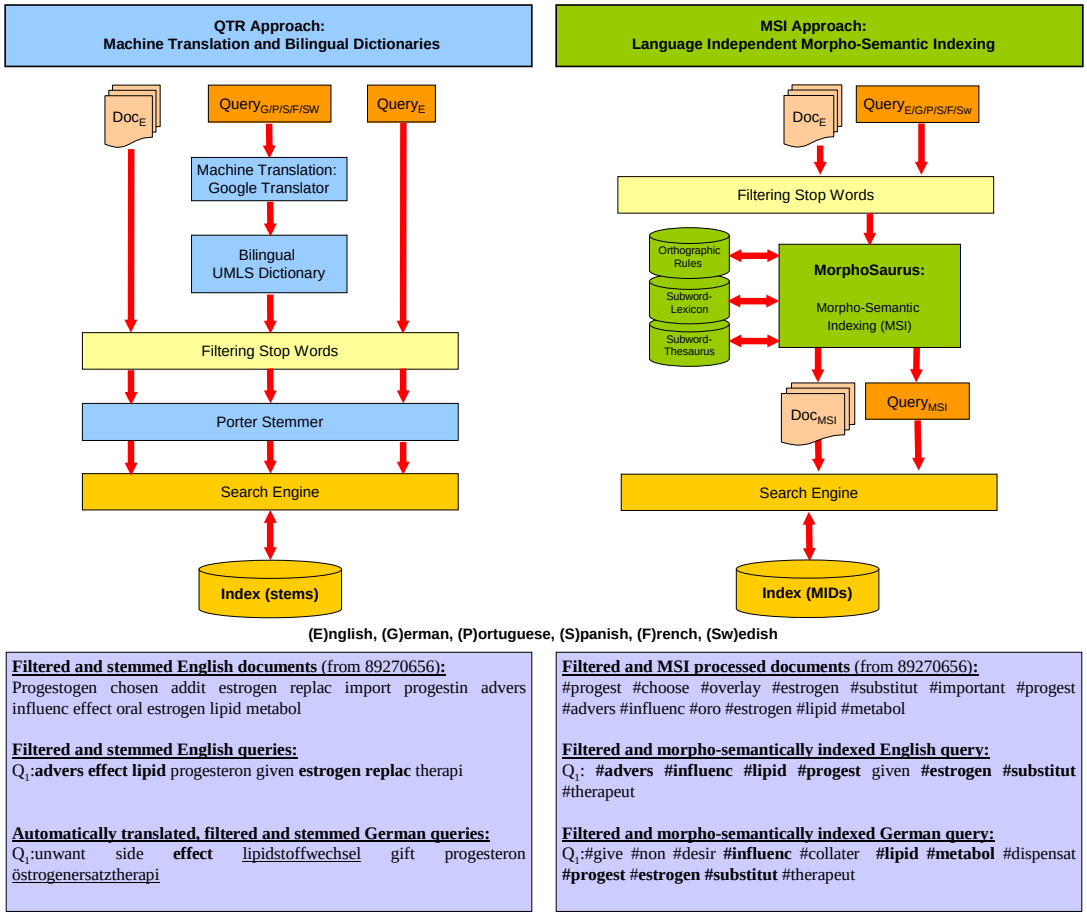

Figure 8.1: Steps for Automatic Translation (Left) and MSI-Indexing (Right)

8.1.3.1 QTR Approach: Machine Translation Based on Bilingual Dictionaries

Machine translation based approaches to CLIR (cf. Oard & Diekema (1998) for an overview) either translate native-language queries into the target language of the document collection to be searched, or otherwise, translate the entire set of documents into each (supported) query language (McCarley, 1999; Rosemblat et al., 2003). Since the latter is naturally a resource intensive task, query translation can be regarded as a standard, and often preferred, experimental procedure in the cross-language retrieval community (Eichmann et al., 1998).

For evaluation, the manually translated queries were re-translated into English

	OHSUMED			IMAGECLEFMED		
Language	Words	GOOGLE	+UMLS	Words	GOOGLE	+UMLS
German	573	496 (86.6%)	522 (91.1%)	136	115 (84.6%)	117 (86.0%)
Portuguese	589	475 (80.6%)	510 (86.6%)	182	158 (86.8%)	161 (88.5%)
Spanish	831	740 (89.1%)	771 (92.8%)	116	94 (81.0%)	99 (85.3%)
French	909	846 (93.1%)	867 (95.4%)	115	106 (92.2%)	107 (93.0%)
Swedish	449	0 (0.0%)	73 (16.3%)	129	0 (0.0%)	9 (7.0%)

Table 8.1: Coverage Statistics for the Automatic Translation of All Query Words Using GOOGLE and UMLS

using the GOOGLE TRANSLATOR.[6] Admittedly, this tool may not be particularly suited to translate medical terminology: considering the OHSUMED collection, 13% of the German, 19% of the Portuguese, 11% of the Spanish and 7% of the French query terms were not translated, while Swedish is not supported at all (cf. Table 8.1, left). Hence, bilingual lexeme dictionaries derived from the UMLS Metathesaurus were used additionally.[7] If no English correspondence could be found, the terms were left untranslated.

Just as in the baseline condition, the stop words were removed from both the documents and the automatically translated queries and potential suffixes were stripped off. The left side of Figure 8.1 visualizes this approach which is referred to as QTR (query translation).

8.1.3.2 MSI-Approach: Language Independent Morpho-Semantic Indexing

As an alternative to QTR, the approach which is based on the morpho-semantic normalization procedures was probed, as introduced in Section 3.2. Unlike QTR, the indexing of documents *and* queries using MSI yields a language-independent,

[6] `http://www.google.de/language_tools`

[7] In contradistinction to the UMLS-derived parallel corpora described in Section 5.2, only word-to-word translations are considered here.

semantically normalized index format. The right side of Figure 8.1 illustrates the basic computation steps for MSI.

8.1.4 Search Engine

For an unbiased evaluation, several experiments were run with LUCENE (Gospodnetic & Hatcher, 2004),[8] a freely available open-source search engine which combines Boolean searching with a sophisticated ranking model based on TF-IDF (Salton & Buckley, 1988). Beside its ranking formula, which achieves results that even can outperform advanced vector retrieval systems (Tellex et al., 2003), this search engine has another advantage: it supports a rich query language like multi-field search, including more than ten different query operators.

In previous experiments, *coordination matching* was used combined with proximity search, which allows to find words within a specified window size. For example, given the query *"NEAR(talar fracture,3)"*, documents are found which contain the words *"talar"* and *"fracture"* within three words distance to each other. It additionally allows word swaps (e.g., *"fracture of the talar bone"*, *"talar bone fracture"*). Evidence has been found that this feature increases the retrieval performance in any scenario, including the baseline condition (Hahn et al., 2004a; Markó et al., 2005c). Especially the effect of considering a window of three items significantly increases the score of clustered matches. This becomes particularly important in the segmentation of complex word forms. Otherwise, a document containing *"append⊕ectomy"* and *"thyroid⊕itis"*, and another one containing *"append⊕ic⊕itis"* and *"thyroid⊕ectomy"* become indistinguishable after segmentation. LUCENE supports proximity search, too. However, the effect on the retrieval performance is counterbalanced by the TF-IDF ranking model so that no improvements can be observed any longer. Therefore, the experiments in the following were performed without using the adjacency constraint.

The preprocessing of documents and queries with the morpho-semantic normalization procedures can generally be adapted to any alternative search engine archi-

[8] `http://jakarta.apache.org/lucene/docs/index.html`

tecture, including simple search on relational databases or based on sophisticated vector space models, a prominent example being the SMART system (Salton, 1971).

8.1.5 Experimental Conditions

Three different basic test conditions can now be distinguished for the retrieval experiments:

- **BASELINE**: The baseline of the experiments is given by the OHSUMED and IMAGECLEFMED corpora both in terms of their Porter-stemmed English queries, as well as their Porter-stemmed (English) document collection.

- **QTR**: In this condition, German, Portuguese, Spanish, French and Swedish queries are automatically translated into English ones (using the GOOGLE TRANSLATOR and the UMLS Metathesaurus), which are Porter-stemmed after the translation. These queries are evaluated on the Porter-stemmed OHSUMED and IMAGECLEFMED document collections.

- **MSI**: This condition stands for the automatic transformation of the German, Portuguese, Spanish, French, and Swedish queries into the language-independent MSI interlingua (plus lexical remainders). The entire OHSUMED and IMAGECLEFMED document collections are also submitted to the MSI procedure. Finally, the MSI-coded queries are evaluated on the MSI-coded corpora, both at an interlingual representation level. In this scenario, four different categories can be further discriminated:

  - **MSI-core**: The experiments were run incorporating neither the acronym module (Section 6), nor the disambiguation module (Section 7).

  - **MSI-D**: The experiments were run incorporating the disambiguation module, but without the acronym module.

  - **MSI-full**: The experiments were run incorporating both the disambiguation and acronym module.

8.1.6 Measurements

Several measurements were taken in comparing the performance of QTR and the different MSI scenarios. The first one is the average of the precision values at all eleven standard recall points (0.0, 0.1, 0.2, ..., 1.0). Furthermore, the average at the top two recall points (0.0 and 0.1) were calculated. While this data was computed with consideration of the first 200 documents under each condition, the exact precision scores for the top five and top 20 ranked documents were also taken into account.

8.2 OHSUMED Results

Considering the different test conditions and languages , Table 8.2 contains the exact numbers (best results for each language marked bold), and Figures 8.2 and 8.3 the corresponding visualizations of the results.

As depicted in Table 8.2 (first Row), the English-English baseline performs with an 11pt average of 0.19 (Column 3). For English, the experiment was also run using the original representations of queries and documents (without stemming and stop word elimination). As can be seen from the data, stemming is beneficial, with an average gain of 10 percentage points (second row). Running the experiment by MSI-indexing the original OHSUMED corpus, the baseline condition can be exceeded up to 116% for English. Additionally using the disambiguation module, 100% of the baseline up to 121% is reached for English, German, and Portuguese. For the other languages considered, this scenario (MSI-D) also yields best results, ranging from 79% (French and Swedish) to 84% (Spanish) of the baseline. The incorporation of the acronym resolution module does not give any additional benefits. Rather than this, that scenario almost performs as good as running MORPHOSAURUS without disambiguation (MSI-core). On the other hand, the QTR approach scores far lower than any MSI condition, reaching 37% of the baseline for Swedish and a maximum of 63% for Spanish. This results in 21 percentage points difference for Spanish up to 53 percentage points for German (QTR compared to MSI-D).

The uneven investment of effort in constructing the different lexicons (mainly

Language	Condition	11pt	top 2 pt	top 5	top 20
English	BASE	.19	.42	.39	.27
English	Original	.17 (89.5)	.36 (85.7)	.36 (92.3)	.25 (92.6)
	MSI-core	.22 (115.8)	.47 (111.9)	.42 **(107.7)**	.29 (107.4)
	MSI-D	.23 **(121.1)**	.48 **(114.3)**	.42 **(107.7)**	.31 **(114.8)**
	MSI-full	.22 (115.8)	.46 (109.5)	.41 (105.1)	.29 (107.4)
German	QTR	.11 (57.9)	.25 (59.5)	.22 (56.4)	.17 (63.0)
	MSI-core	.20 (105.3)	.41 **(97.6)**	.36 (92.3)	.27 (100.0)
	MSI-D	.21 **(110.5)**	.41 **(97.6)**	.37 **(94.9)**	.28 **(103.7)**
	MSI-full	.20 (105.3)	.40 (95.2)	.35 (89.7)	.27 (100.0)
Portuguese	QTR	.11 (57.9)	.24 (57.1)	.21 (53.8)	.15 (55.6)
	MSI-core	.17 (89.5)	.37 (88.1)	.36 (92.3)	.23 (85.2)
	MSI-D	.19 **(100.0)**	.39 **(92.9)**	.37 **(94.9)**	.25 **(92.6)**
	MSI-full	.18 (94.7)	.38 (90.5)	.36 (92.3)	.25 **(92.6)**
Spanish	QTR	.12 (63.2)	.25 (59.5)	.23 (59.0)	.16 (59.3)
	MSI-core	.16 **(84.2)**	.36 **(85.7)**	.32 **(82.1)**	.22 **(81.5)**
	MSI-D	.16 **(84.2)**	.36 **(85.7)**	.32 **(82.1)**	.22 **(81.5)**
	MSI-full	.16 **(84.2)**	.35 (83.3)	.32 **(82.1)**	.22 **(81.5)**
French	QTR	.10 (52.6)	.23 (54.8)	.20 (51.3)	.16 (59.3)
	MSI-core	.12 (63.2)	.23 (54.8)	.23 (59.0)	.15 (55.6)
	MSI-D	.15 **(78.9)**	.31 **(73.8)**	.30 **(76.9)**	.20 **(74.1)**
	MSI-full	.14 (73.7)	.30 (71.4)	.28 (71.8)	.20 **(74.1)**
Swedish	QTR	.07 (36.8)	.16 (38.1)	.11 (28.2)	.10 (37.0)
	MSI-core	.15 **(78.9)**	.31 **(73.8)**	.29 **(74.4)**	.22 **(81.5)**
	MSI-D	.15 **(78.9)**	.30 (71.4)	.29 **(74.4)**	.20 (74.1)
	MSI-full	.14 (73.7)	.29 (69.0)	.28 (71.8)	.20 (74.1)
Average	QTR	.12 (63.2)	.26 (61.9)	.23 (59.0)	.17 (63.0)
	MSI-core	.17 (89.5)	.36 (85.7)	.33 (84.6)	.23 (85.2)
	MSI-D	.18 **(94.7)**	.37 **(88.1)**	.35 **(89.7)**	.24 **(88.9)**
	MSI-full	.17 (89.5)	.36 (85.7)	.33 (84.6)	.24 **(88.9)**

Table 8.2: Precision for the OHSUMED Collection (% of Baseline in Brackets, Best Results Marked Bold)

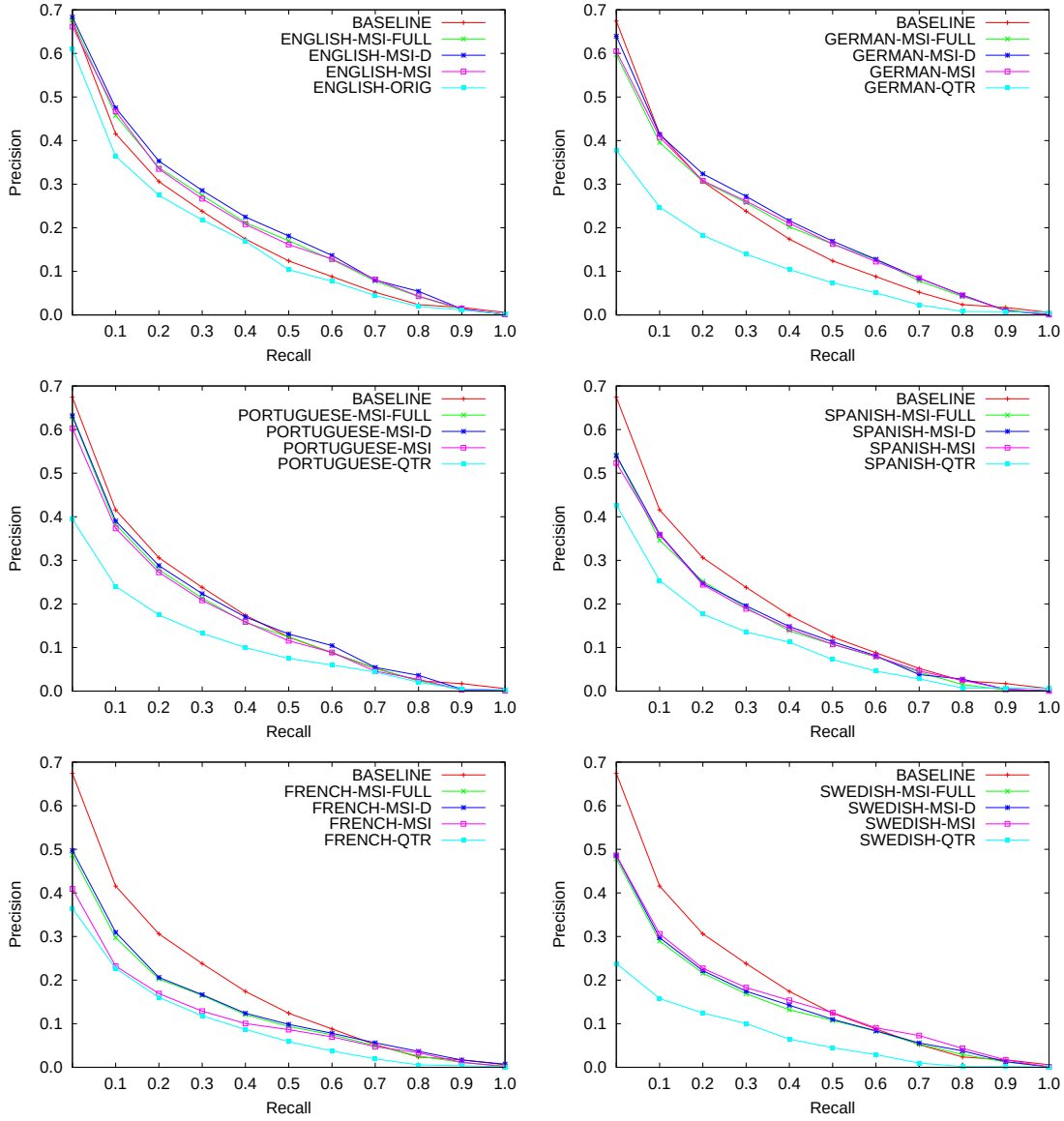

Figure 8.2: Average Precision/Recall Graphs for the OHSUMED Collection

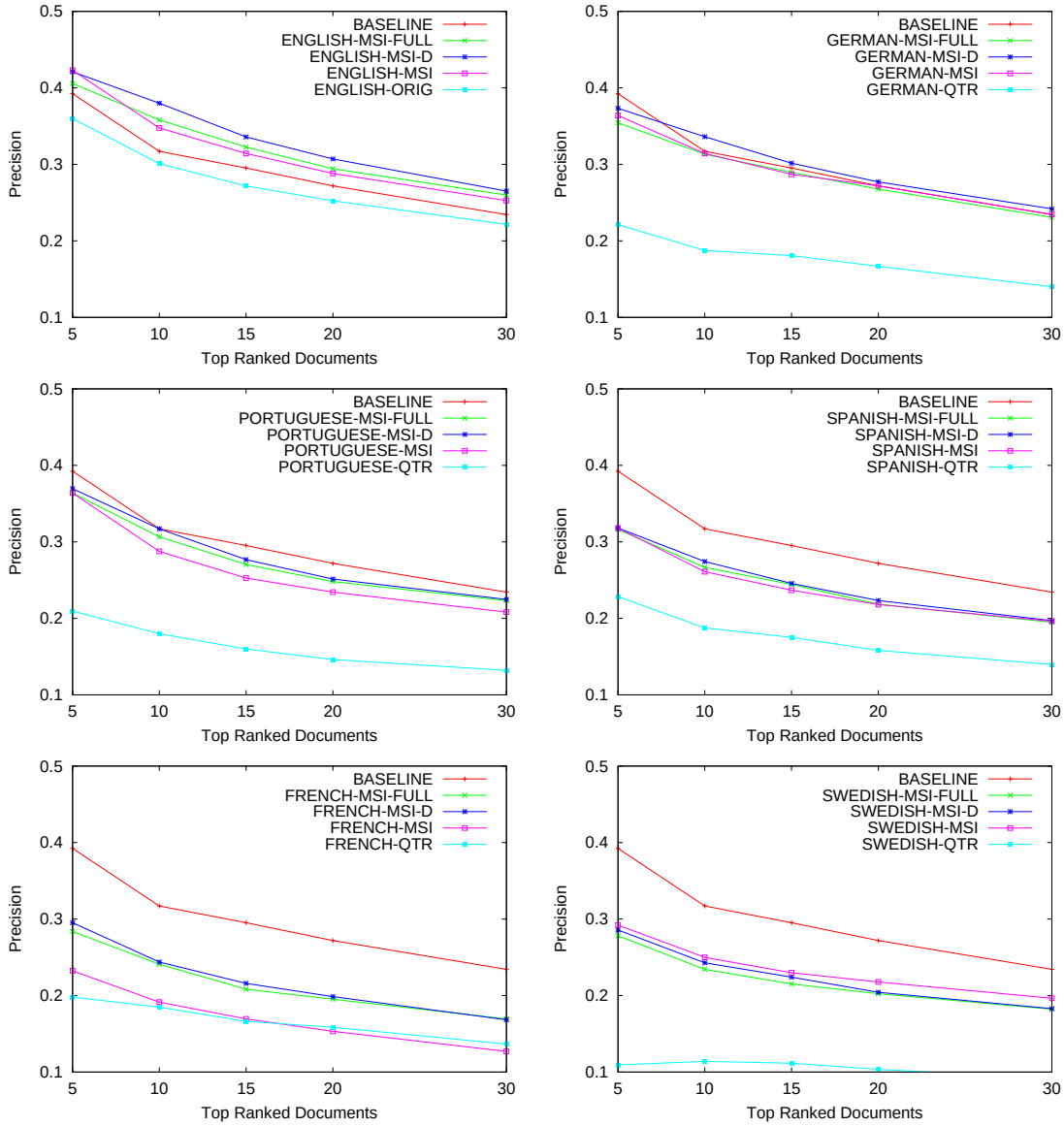

Figure 8.3: Exact Precision Graphs for the OHSUMED Collection

automatically acquired Spanish, French and Swedish entries) is well reflected in the results. In any case, it seems worth noting that at *no* recall point QTR values were higher than MSI values. Hence, the latter throughout outperforms the former on *all* languages.

Interesting from a realistic retrieval perspective is the average gain on the top two recall points. In Table 8.2 (column four). Just as for the 11pt average, MSI-core and MSI-D perform best with values ranging between 74% (Swedish) up to 98% (German) of the baseline. In a monolingual (English) retrieval setting, the baseline can even be exceeded by 14%.

However, there may be considerable variation regarding the actual numbers behind these levels of recall. Medical decision-makers under time pressure are often interested in a few top-ranked documents. Thus, the exact precision scores for these documents are the most indicative of the performance of all approaches discussed here. Considering only the top 5 (Column 5) and top 20 (Column 6) ranked documents, precision does not fall below 74% of the baseline for MSI-D. In contrast, QTR does not exceed 59%, which means that MSI-D clearly outperforms QTR in any language condition. Again, focusing on the (English) monolingual retrieval setting, MSI-D gains 15% compared to the baseline.

By averaging over all languages and adding the English baseline condition to the values of the QTR approach for the other languages, query translation has a mean average precision (11pt) of 0.12, thus reaching 63% of the baseline. MSI-core and MSI-full achieve 90% while MSI-D performs best with 95%. Obviously, the incorporation of the acronym resolution module does not lead to a further benefit compared to MSI-D. The reason for this lies in one peculiarity of the text genre considered: In general, long forms of acronyms which are relevant to a particular MEDLINE document are given in the corresponding abstract. Therefore, mappings between document and query terms are available both for acronyms and respective acronym definitions. Regarding other document types, e.g. clinical findings or discharge summaries, it is likely that no corresponding definitions are given (see next section). Another reason is that additional noise is entered to the data, since highly ambiguous acronyms have to be disambiguated correctly. Regarding the 11pt

average, the benefit of the possibility to the search for acronyms across languages is negated by the uncertainty resulting from resolving acronyms. However, considering only the top 20 ranked documents, the acronym module does not hamper cross-lingual retrieval, though it does not yield any additional boost.

8.3 ImageCLEFMed Results

Table 8.3 depicts the results for the ImageCLEFMed corpus. Since there are only 30 queries in this collection, the graphs in Figures 8.4 and 8.5 which show the corresponding visualizations of the data, are less smooth.

Unlike the Ohsumed collection, in which documents consist of coherent texts (MEDLINE abstracts), ImageCLEFMed contains short captions of medical images, often only consisting of noun phrases with many acronyms. This might be the reason why the overall-performance is not comparable to Ohsumed for all scenarios, including the baseline condition. This is, in particular, witnessed by the small gain of only 6% using stemming on the English monolingual condition.

Though the baseline can not be exceeded in any scenario, the advantage of MSI compared to QTR is throughout observable. While, on the average, QTR yields 71% of the baseline regarding 11pt average, incorporating the acronym resolution module in the MSI condition performs best with 82%. MSI-D still reaches 77%, while MSI-core and QTR perform equally well, since the Google translator performs surprisingly good on non-English queries, leaving the relatively high amount of acronyms unchanged (just as MSI-core and MSI-D). Only MSI-full is capable of realizing the language transfer of acronyms, e.g. English *"CT - computed tomography"* to Portuguese *"TC - tomografia computadorizada"* or English *"MRI - magnetic resonance imaging"* to French *"IRM - imagerie résonance magnétique"*. Therefore, more relevant documents can be found in the collections. On the other hand, considering only a few top ranked documents, no difference between MSI-full and MSI-D is observable (top 5). For the first 20 retrieved documents, MSI-D even performs better, but keeping in mind that ImageCLEFMed provides a relatively small sample of only 30 queries. This is also the reason why the best cross-lingual

Language	**Condition**	**11pt**	**top 2 pt**	**top 5**	**top 20**
English	BASE	.17	.36	.48	.36
English	Original	.16 **(94.1)**	.33 **(91.7)**	.39 (81.3)	.30 (83.3)
	MSI-core	.15 (88.2)	.28 (77.8)	.44 **(91.7)**	.35 **(97.2)**
	MSI-D	.16 **(94.1)**	.31 (86.1)	.44 **(91.7)**	.35 **(97.2)**
	MSI-full	.15 (88.2)	.29 (80.6)	.40 (83.3)	.30 (83.3)
German	QTR	.10 (58.8)	.21 (58.3)	.28 (58.3)	.22 (61.1)
	MSI-core	.13 (76.5)	.27 (75.0)	.43 (89.6)	.33 (91.7)
	MSI-D	.13 (76.5)	.28 (77.8)	.45 **(93.8)**	.34 **(94.4)**
	MSI-full	.14 **(82.4)**	.29 **(80.6)**	.44 (91.7)	.33 (91.7)
Portuguese	QTR	.13 **(76.5)**	.24 (66.7)	.31 (64.6)	.22 (61.1)
	MSI-core	.10 (58.8)	.25 (69.4)	.37 (77.1)	.31 **(86.1)**
	MSI-D	.12 (70.6)	.27 **(75.0)**	.44 **(91.7)**	.30 (83.3)
	MSI-full	.13 **(76.5)**	.26 (72.2)	.44 **(91.7)**	.30 (83.3)
Spanish	QTR	.13 (76.5)	.27 (75.0)	.32 (66.7)	.22 (61.1)
	MSI-core	.13 (76.5)	.28 (77.8)	.38 (79.2)	.32 (88.9)
	MSI-D	.13 (76.5)	.28 (77.8)	.39 (81.3)	.32 (88.9)
	MSI-full	.14 **(82.4)**	.29 **(80.6)**	.42 **(87.5)**	.33 **(91.7)**
French	QTR	.10 (58.8)	.18 (50.0)	.25 (52.1)	.18 (50.0)
	MSI-core	.11 (64.7)	.23 (63.9)	.34 (70.8)	.30 (83.3)
	MSI-D	.12 **(70.6)**	.25 **(69.4)**	.37 **(77.1)**	.32 **(88.9)**
	MSI-full	.12 **(70.6)**	.23 (63.9)	.35 (72.9)	.30 (83.3)
Swedish	QTR	.06 (35.3)	.09 (25.0)	.16 (33.3)	.08 (22.2)
	MSI-core	.12 (70.6)	.27 (75.0)	.36 (75.0)	.32 (88.9)
	MSI-D	.13 (76.5)	.27 (75.0)	.39 **(81.3)**	.34 **(94.4)**
	MSI-full	.14 **(82.4)**	.29 **(80.6)**	.39 **(81.3)**	.30 (83.3)
Average	QTR	.12 (70.6)	.23 (63.9)	.3 (62.5)	.21 (58.3)
	MSI-core	.12 (70.6)	.26 (72.2)	.39 (81.3)	.32 (88.9)
	MSI-D	.13 (76.5)	.28 **(77.8)**	.41 **(85.4)**	.33 **(91.7)**
	MSI-full	.14 **(82.4)**	.28 **(77.8)**	.41 **(85.4)**	.31 (86.1)

Table 8.3: Precision for the ImageCLEFMed Collection (% of Baseline in Brackets, Best Results Marked Bold)

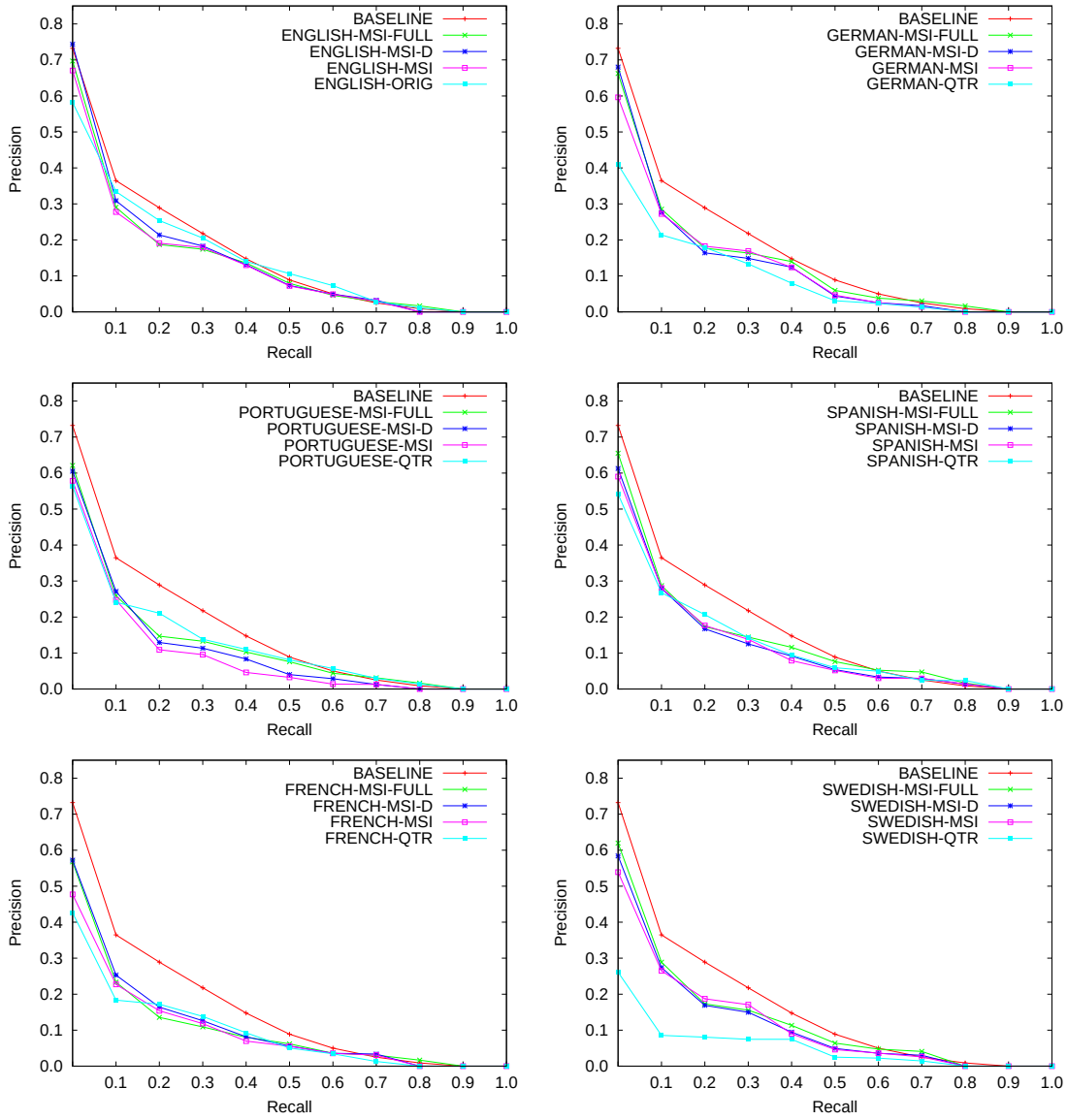

Figure 8.4: Average Precision/Recall Graphs for the ImageCLEFMed Collection

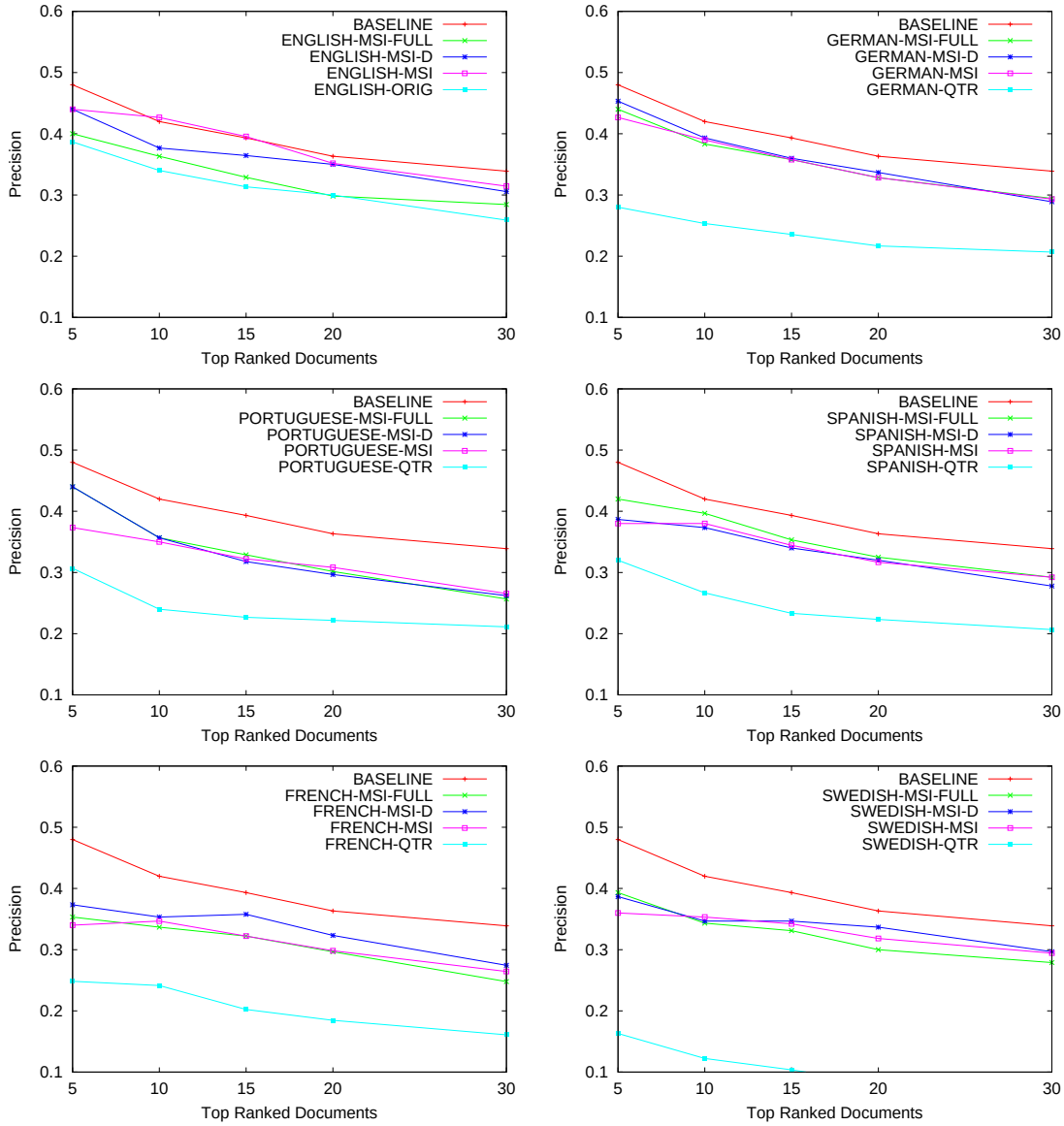

Figure 8.5: Exact Precision Graphs for the ImageCLEFMed Collection

results (besides German) are achieved for Spanish and Swedish (82% of the baseline considering 11pt average), though Spanish and Swedish are less covered by the lexicons underlying MorphoSaurus than Portuguese (77%).

Summarizing, if solely using ImageCLEFMed for the evaluation of (Cross-Language) Information Retrieval systems, this is of limited value only. However, since the collection focuses on an important medical subdiscipline (medical imaging and picture archiving systems) and results are in-line with those using the Ohsumed collection, additional evidence for the excellent performance of MorphoSaurus in a Cross-Language Information Retrieval setting is available.

8.4 Discussion

After more than a decade of intensive research, Cross-Language Information Retrieval (CLIR) has produced considerable achievements (Gey et al., 2002). From a methodological point of view, the field of CLIR is divided into dictionary-based *vs.* corpus-based approaches (Oard & Diekema, 1998). Since corpus-based approaches depend on the availability of large parallel corpora, which is mostly not the case for technical sublanguages, most efforts in CLIR are centered around either query translation, expansion and structuring, or document translation (Rosemblat et al., 2003). McCarley (1999) reports on a translation model, which incorporates both query and document translation and outperforms either translation direction. A more recent strategy for machine translation based CLIR is the use of commercial software (Savoy, 2003b), which usually provides only poor support of technical sublanguages. For medical terminology and other sublanguages, non-specialized multilingual lexicons (based on WordNet) also offer limited support only (Gonzalo et al., 1999).

The success of dictionary-based CLIR largely depends on the coverage of the lexicon, tools for conflating morphological variants, phrase and proper name recognition as well as word sense disambiguation (Pirkola et al., 2001). Within the MorphoSaurus system, the lexical coverage is optimized by limiting the lexicon to semantically relevant subwords of the medical domain. This also helps in dealing

with morphological variation, including single-word decomposition. Since the latter is a very common phenomenon in medical terminology, this partially explains the poor results for German in the SAPHIRE medical text retrieval system which used the UMLS Metathesaurus for semantic indexing (Hersh & Donohoe, 1998).

The UMLS, together with WORDNET, is also the lexical basis of the approach pursued by the MUCHMORE project (Volk et al., 2002). Here, concept mapping occurs after various steps of linguistic pre-processing, including lemmatization (also cf. Rosemblat et al. (2003) and Rosemblat & Graham (2006)). Although good results are communicated, these are not comparable to those presented here because the authors use home-grown document and query collections and diverge in the construction of their baseline.

Chen (2002) proved decompounding for German and Dutch to be effective in monolingual and bilingual retrieval. He uses bilingual lexicons and a probabilistic decomposition strategy for which the mean precision increase ranges from 8,4% to 11,46% for German or French to English, respectively.

Stemming is beneficial in a monolingual scenario, as reported by Braschler & Ripplinger (2004), even when a simple approach is used (also cf. the first two rows in Tables 8.2 and 8.3). Carefully designed decomposition remarkably boosts performance. Applying stemming to queries and documents yields a performance gain in mean average precision of up to 23%. Decomposition contributes even more to performance improvement than stemming with values up to 34% for short queries. The same stemming algorithm was in use in a multilingual scenario (Braschler & Schäuble, 2000; Braschler et al., 2003) in which the authors demonstrated the advantages of decompounding in CLIR. Other monolingual settings reporting a performance gain when applying linguistic analysis are described in the work of Tomlinson (2001) and Moulinier et al. (2001).

Eichmann et al. (1998) report on cross-language experiments for French and Spanish using the same test collection as used here (OHSUMED), and the UMLS Metathesaurus for query translation, achieving 71% of their baseline for Spanish and 61 % for French (contrasted to 84% and 79% for MSI-D, respectively). With the SMART-style vector space engine they employ (Salton, 1971), their overall 11pt

performance (0.24) is far above the one determined here (0.18). On the other hand, when focusing on exact precision scores that have more explanatory power when thinking of a real world user scenario, the MorphoSaurus approach turns out to be more advantageous. Eichmann et al. report precision values of 0.23 for Spanish (0.17 for French) for the top 5 ranked documents and 0.21 (0.14, respectively) for the top 10. Compared to these scores, the MSI approach (involving disambiguation) reaches 0.32 for Spanish (0.30 for French) for the top 5 ranked documents (cf. Table 8.2). Even when discarding disambiguation the MSI approach still outperforms the compared system. Since query translation via the UMLS Metathesaurus was adopted in the work of Eichmann et al., it is not surprising that the QTR scenario thoroughly yields comparable results.

Chapter 9

Cross-Language Information Retrieval on the Web

Despite the wide range of CLIR applications that have been developed in the recent years, only few have been adopted by large Web search engines, online newspapers or information services. The reason for this is, amongst others, that different genres exist in which CLIR may be applied and in which the CLIR techniques are not yet sophisticated enough (Oard, 2002; Gey & Peters, 2005). This holds specifically for scientific and technical literature.

In the previous chapter, it has been shown that subword decomposition of both documents and queries can significantly improve the performance of both intralingual and cross-lingual document retrieval in the medical domain. Having only limited resources, this approach is only suitable for "closed" document collections which can be stored locally. Now, in order to expand search facilities to the Web, morphosemantic indexing can be used to manipulate the original queries (and not the documents). The resulting interlingual representation is then the basis for the translation of queries into (a) desired target language(s) in a second step. Afterwards, any Internet search engine can be used to retrieve relevant documents from the Web.

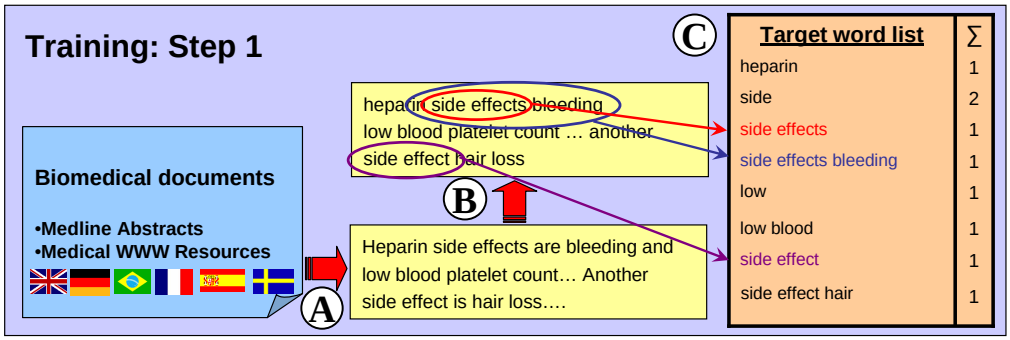

Figure 9.1: Training Target Words for the Translation Process

9.1 Query Translation for Web-CLIR

The query translation process can generally be regarded as a two step procedure. In a preparation phase, large language-specific corpora are divided into sequences of words (n-grams). These sequences are then processed by MorphoSaurus. As a result, a collection of MID sequences associated to sequences of word-n-grams are stored in look-up tables for different target languages. Afterwards, when a query is entered by a user, it is also processed into a set of corresponding MIDs. Based on this representation, different syntactic readings are generated that are divided into blocks of possibly coherent MIDs. These blocks serve as the link to the word-n-grams of the desired target language and are, therefore, compared to the MIDs in the correspondent look-up table. Possible translations are returned as a ranked output list ordered by a frequency score.

9.1.1 Creating Subword Lists

In the preparation phase, large (medical) domain specific corpora in different languages from the Web are used (cf. Figure 9.1, step A), including abstracts from medical journals indexed in MEDLINE (cf. Table 5.3 in Section 5.1.1.1). Stop words are filtered from these resources and characters transferred to lower-case (Figure 9.1, step B). Subsequently, these corpora are tokenized into word-n-grams (henceforth, *target words*, cf. Figure 9.1, step C). By limiting n to values between 1 and 3, lists of surface words, word bigrams and trigrams are obtained. These temporary lists are uniquely sorted, counting the number of occurrences. Table 9.1 lists the number

Language	Surface Words	Bigrams	Trigrams
English	528,585	30,257,162	97,673,610
German	467,909	4,101,444	5,530,952
Portuguese	138,248	3,899,548	7,058,870
Spanish	126,314	2,382,785	3,746,541
French	85,710	1,129,152	1,796,513
Swedish	47,343	423,625	782,648

Table 9.1: Number of Generated Target Words in Different Languages

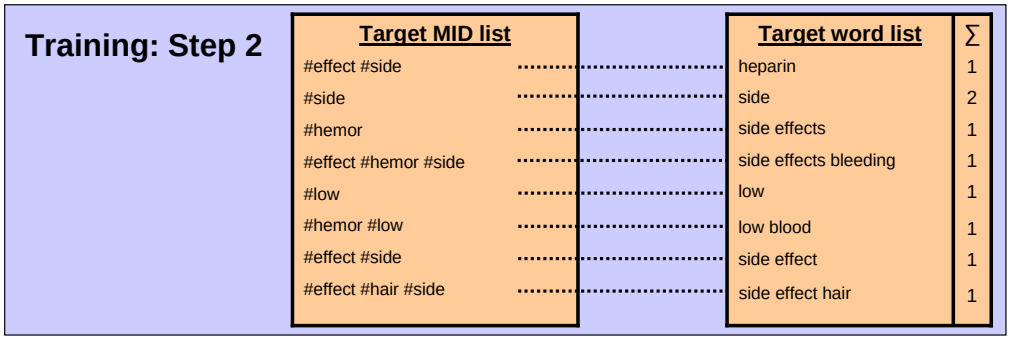

Figure 9.2: Morpho-semantic Normalization of Target Words

of generated word-n-grams for English, German, Portuguese, Spanish, French and Swedish.

The target words are now processed with the morpho-semantic normalization routine which associates each word-n-gram with a sequence of MIDs (Figure 9.2). Then, the resulting language specific *target lists* contain triples of the form (*target words*, *frequency*, *MIDs*). Due to the frequent occurrence of subword permutations between languages (e.g. German *"Bluthochdruck"* (literally *"blood high pressure"*) vs. English *"high blood pressure"* or Swedish *"högt blodtryck"*), bigrams and trigrams on the interlingual MID layer are ordered alphabetically. Table 9.2 shows a small subset of the English *target list*.

Target Words	Frequency	MIDs
...	...	...
side	111,675	#side
side effects	76,366	#effect #side
pancreatitis	9194	#itis #pancreas
heparin	574	#heparin
inflammation pancreas	269	#itis #pancreas
effects asthma	17	#asthma #effect
effects asthma gastric	1	#asthma #effect #gastr
...	...	...

Table 9.2: Extract of the English Target List

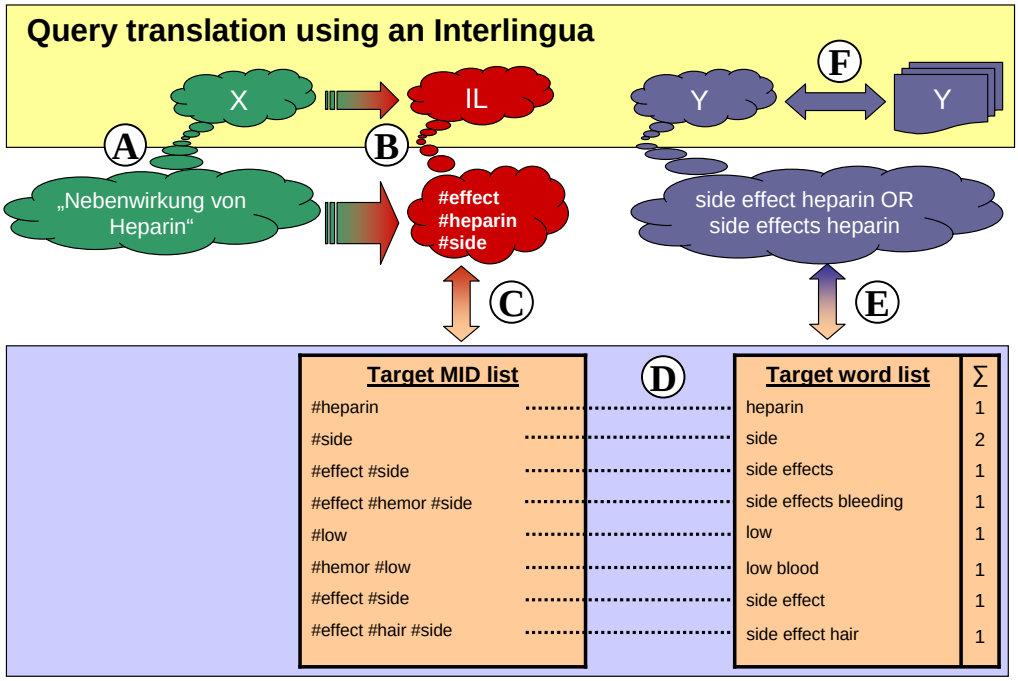

Figure 9.3: Producing Translations: A User Query in Language X is Transformed into the Interlingua IL from which it is Mapped to a Word List in a Specific Target Language Y.

Q_{orig}	*Nebenwirkungen von Heparin*
Q_{MID}	#side #effect #heparin
Partitions	#side #effect #heparin
	#side | #effect #heparin
	#side #effect | #heparin
	#side | #effect | #heparin

Table 9.3: Possible Syntactic Readings for Query Q_{orig}

9.1.2 Producing Translations

When a user query Q_{orig} is sent to the query translation tool (with specified query language and desired target document language), Q_{orig} is transformed to its MID representation Q_{MID} (sketched in Figure 9.3, step A and B). For Q_{MID}, a list of possible syntactic readings is generated by adding or omitting the delimiter sign "|" between each pair of consecutive MIDs. A possible reading in this list is, by definition, called a *partition*, and a partition element (a set of MIDs between two vertical delimiters) forms a *subquery*. Subqueries represent possible coherent units in a query and serve as a base for the subsequent translation step between the interlingua and the target language.

As an example, taking the German Q_{orig} *"Nebenwirkungen von Heparin"* (English: *"side effects of heparin"*), Q_{orig} is transformed to the MID representation Q_{MID} =[#side #effect #heparin]. Subsequently, a list of possible syntactic readings is produced, as depicted in Table 9.3.

After the MIDs of each subquery are ordered alphabetically the subqueries are then matched against the MIDs in the target list (Figure 9.3, step C). The first n_{HIT} hits are returned which represent possible translations for a corresponding subquery. In a following step, all subqueries in the partitions are replaced by their corresponding translations (i.e. *target words*) in the target language, thus obtaining a list of possible translations for Q_{orig} (Figure 9.3, step D).

The number of possible translations can be high: Having a query with n_{MID} MIDs, $2^{n_{MID}-1}$ partitions are obtained. The number of partitions in relation to

Subquery	Target Word	Frequency
#side #effect #heparin	side effects heparin	14^A
	side effect heparin	13^B
#side #effect	side effects	$76{,}366^C$
	side effect	3,856
#effect #heparin	effect heparin	112
	effects heparin	94
#side	side	111,675
	lateral	24,632
#effect	effects	353,466
	effect	244,994
#heparin	heparin	$12{,}365^C$
	heparins	570

Table 9.4: Subqueries and their Two most Frequent Matches in the Target List

their number of subqueries (n_{SQ}) follows a binomial distribution. Thus, a query with n_{MID} MIDs and n_{SQ} subqueries has $\binom{n_{MID}-1}{n_{SQ}-1}$ partitions. The maximum number of translations, n_{TR}, is computed as follows:

$$n_{TR} = \sum_{i=1}^{n_{SQ}} \binom{n_{MID}-1}{n_{SQ}-1} * n_{HIT}^{n_{SQ}}, \text{ with}$$

n_{SQ}: the number of *subqueries* in a *partition*
n_{MID}: the number of MIDs in a *partition*
n_{HIT}: the number of hits in the *target list*

As an example, *"Nebenwirkungen von Heparin"* translates to [#side #effect #heparin] on the interlingual layer($n_{MID} = 3$). Allowing four hits to be returned for each subquery ($n_{HIT} = 4$), a maximum number of $n_{TR} = 1*4^1 + 2*4^2 + 1*4^3 = 100$ possible translations is obtained (also cf. Table 9.5).

9.1.3 Ranking of Translations

Given this amount of possible translations there is a need for a reasonable ranking algorithm. For this purpose, the length of subqueries and the frequencies of occurrence of the target words (counted in the training phase, cf. Section 9.1.1) serve as a measure for the lexical importance to compute a ranking score of the translation candidates.

Considering the partitions listed in Table 9.3, results from matching their *subqueries* against the English target list are depicted in Table 9.4 (here, two hits are returned for each subquery, $n_{HIT} = 2$) .

Taking this frequency data as a base, the ranking algorithm can now be described as follows:

1. Use all translations as candidates that correspond to the partition containing exactly one subquery ($n_{SQ} = 1$).

2. Having n_{SQ} subqueries in a partition, denoted by $SQ_{1..n_{SQ}}$, and $|SQ_j|$ denoting the number of MIDs in a particular subquery ($1 \leq j \leq n_{SQ}$) and $freq_{tw_j}$ denoting the frequency of occurrence of target words in the target list which are associated to SQ_j, rank all i translation candidates according to their score that is computed as follows:

$$score_i = \sqrt[n_{SQ}]{\prod_{j=1}^{n_{SQ}} freq_{tw_j}^{|SQ_j|}}$$

3. If no candidates can be found in Step 1 or more translations are required, increase the number of subqueries in a partition by one (n_{SQ} += 1). Again, find all candidates that correspond to the partitions subqueries and repeat Step 2.

Table 9.5 shows possible translations for each partition of the example. According to Step 1 of the ranking algorithm, the partition [#side #effect #heparin] is considered firstly. The *score* of the corresponding translations is computed by 14^3 (cf. Table 9.4, ^{A}) or 13^3 (cf. Table 9.4, ^{B}), respectively. Afterwards, all translations for partitions covering two subqueries ($n_{SQ} = 2$) are considered. The frequency score

Subquery	Target Word	Score
#side #effect #heparin	side effect heparin	2,744[A]
	side effects heparin	2,197[B]
#side #effect | #heparin	side effects heparin	8,491,748[C]
	side effects heparins	1,823,213
	side effect heparin	428,779
	side effect heparins	92,060
#side | #effect #heparin	side effect heparin	37,427
	side effects heparin	31,412
	lateral effect heparin	17,577
	lateral effects heparin	14,752
#side | #effect | #heparin	side effects heparin	78,734
	side effects heparins	28,230
	side effect heparin	69,678
	side effect heparins	24,984
	lateral effects heparin	47,571
	lateral effects heparins	17,057
	lateral effect heparin	42,100
	lateral effect heparins	15,095

Table 9.5: Subqueries, Query Translations and their Scores

for the translation *"side effects heparin"*, resulting from the partition [#side #effect | #heparin] (marked with ^{C} in Table 9.4), is exemplarily computed as follows:

$$score^C = \sqrt[2]{76366^2 * 12365} = 8,491,748$$

After removing duplicates, a ranked list of possible translations is generated, as depicted in Table 9.6. Erroneous translations (5-8) are ranked at the bottom of the list. Taking the first n entries in the translation set allows to automatically construct a disjunctive query, as depicted in Figure 9.3 (step E), which can then be sent to any Web search engine for retrieving documents in a specific target language (Figure 9.3, step F).

Rank	Translation
1.	side effects heparin
2.	side effects heparins
3.	side effect heparin
4.	side effect heparins
5.	lateral effects heparin
6.	lateral effect heparin
7.	lateral effects heparins
8.	lateral effect heparins

Table 9.6: Ranked List of Possible Translations of the German Phrase "Nebenwirkungen von Heparin"

9.2 Interface to a Web Search Engine

To demonstrate the potential of translating queries using MorphoSaurus as a basis (Daumke et al., 2005a; 2005b), an interface to the most popular Web search engine has been created.[1] Screenshots of this application are shown in Figure 9.4 (Web search) and Figure 9.5 (search restricted to PubMed, the interface to the MEDLINE database maintained by the U.S. National Library of Medicine). In the prototype version, the user can choose (amongst other parameters) the maximum number of hits per subquery (n_{HIT}), as well as the number of translations to be sent to the search engine (n_{TR}). Intrinsically, all possible translations could be sent to this search engine combined with an OR operator. Since the interface of the search engine limits the number of tokens (i.e. words or operators) to a maximum of ten, each translation is sent to the search engine separately. The subsequent merging algorithm of the different search engine results ranks those items at the top that are found in more than one search run. All others are added at the bottom of the result list in a simple fashion, adding the best hit of each result list iteratively until all hits are processed.

[1] `http://www.google.com`

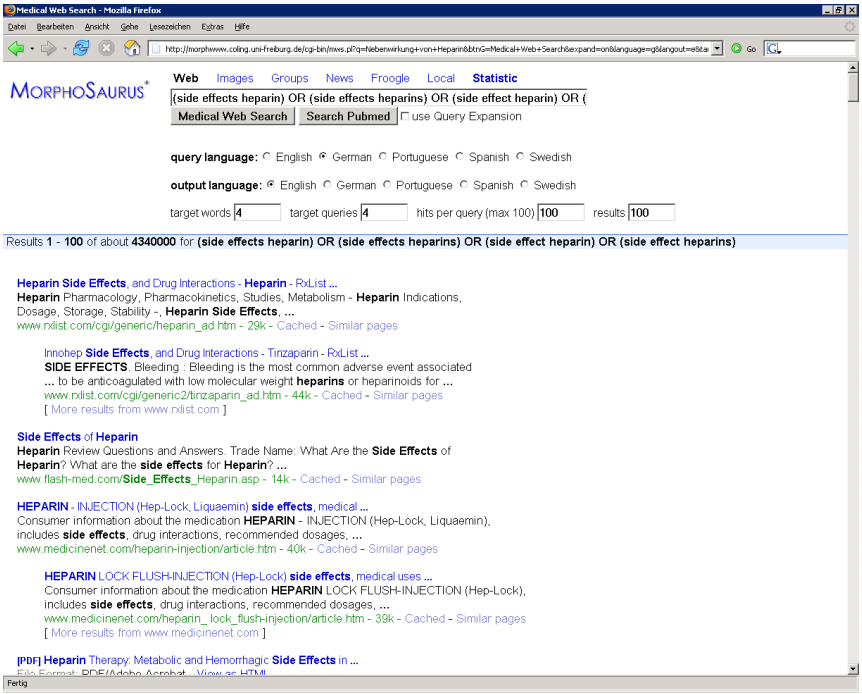

Figure 9.4: Subword-based CLIR on the Web

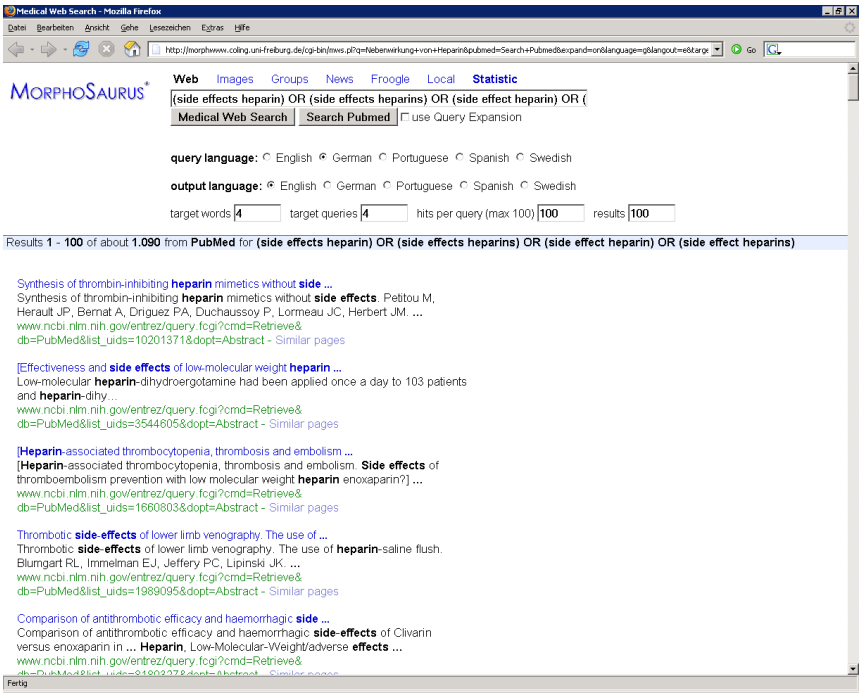

Figure 9.5: Subword-based CLIR on NLM's PubMed

9.3 Evaluation

Since it is not feasible to evaluate Information Retrieval on the Web in terms of precision and recall, the same test sets as used for evaluating MorphoSaurus by matching document terms and query terms at the interlingual layer, Ohsumed and ImageCLEFMed (cf. Section 8.1), are also used in the following. Retrieval results for query translation based on cross-lingual, morpho-semantic indexing of subwords (MSI-QTR), is compared to the monolingual baseline (BASE), where English queries are matched against English documents, and direct query translation (QTR), as described in Section 8.1.3.1. Furthermore, the contribution of query expansion, i.e. using five or ten possible translations of an original query is analyzed in detail. Again, Lucene was used as the underlying retrieval system.

Most Web search engines such as Google or Yahoo! do not incorporate stemming, or only to some extent. The reason for this is that search engine providers focus on precision rather than recall. Their Web crawlers index more than 8 billion pages. So it is likely that enough relevant pages can be found for a single query, and stemming would decrease precision noticeably. Therefore, the baseline condition for the experiments is based on unstemmed original documents and queries (cf. Row 2 in Tables 8.2 and 8.3 in the previous chapter for the results from stemmed texts).

9.4 Ohsumed Results

As shown in Table 9.7 (Column 3), regarding the 11pt average values, precision values for the QTR approach range between 0.07 (French) and 0.10 (Portuguese). This means that QTR values constitute between 41% and 59% of the baseline condition. In contrast, when applying the MSI-QTR method allowing only one translation (MSI-QTR-1), precision results vary from 0.09 (Spanish) to 0.16 (German). Consequently, the relative performance is between 53% (Spanish) and 94% (German), which makes a difference of up to 41 percentage points. For English, applying MSI-QTR-1 means that the query is replaced by a similar one that is more likely to be found in the training data. Regarding the mean average precision, no difference to the original query can be observed in this scenario.

Language	Condition	11pt	top 2 pt	top 5	top 20
English	BASE	.17	.36	.36	.25
English	MSI-QTR-1	.17 (100.0)	.34 (94.4)	.34 (94.4)	.23 (92.0)
	MSI-QTR-5	.19 **(111.8)**	.39 **(108.3)**	.38 **(105.6)**	.26 **(104.0)**
	MSI-QTR-10	.19 **(111.8)**	.39 **(108.3)**	.37 (102.8)	.26 **(104.0)**
German	QTR	.09 (52.9)	.20 (55.6)	.18 (50.0)	.14 (56.0)
	MSI-QTR-1	.16 (94.1)	.33 (91.7)	.33 (91.7)	.22 (88.0)
	MSI-QTR-5	.17 (100.0)	.36 **(100.0)**	.34 **(94.4)**	.24 **(96.0)**
	MSI-QTR-10	.18 **(105.9)**	.36 **(100.0)**	.34 **(94.4)**	.24 **(96.0)**
Portuguese	QTR	.10 (58.8)	.19 (52.8)	.18 (50.0)	.12 (48.0)
	MSI-QTR-1	.13 (76.5)	.29 (80.6)	.28 (77.8)	.19 (76.0)
	MSI-QTR-5	.15 **(88.2)**	.33 **(91.7)**	.31 **(86.1)**	.22 **(88.0)**
	MSI-QTR-10	.15 **(88.2)**	.33 **(91.7)**	.30 (83.3)	.21 (84.0)
Spanish	QTR	.09 **(52.9)**	.19 **(52.8)**	.19 **(52.8)**	.13 **(52.0)**
	MSI-QTR-1	.09 **(52.9)**	.18 (50.0)	.17 (47.2)	.13 **(52.0)**
	MSI-QTR-5	.09 **(52.9)**	.18 (50.0)	.17 (47.2)	.13 **(52.0)**
	MSI-QTR-10	.09 **(52.9)**	.19 **(52.8)**	.18 (50.0)	.12 (48.0)
French	QTR	.07 (41.2)	.16 (44.4)	.13 (36.1)	.11 (44.0)
	MSI-QTR-1	.10 (58.8)	.21 (58.3)	.20 (55.6)	.16 **(64.0)**
	MSI-QTR-5	.10 (58.8)	.23 **(63.9)**	.22 **(61.1)**	.16 **(64.0)**
	MSI-QTR-10	.11 **(64.7)**	.22 (61.1)	.20 (55.6)	.16 **(64.0)**
Swedish	QTR	.09 (52.9)	.17 (47.2)	.14 (38.9)	.12 (48.0)
	MSI-QTR-1	.10 (58.8)	.22 (61.1)	.21 (58.3)	.14 (56.0)
	MSI-QTR-5	.11 **(64.7)**	.24 **(66.7)**	.22 **(61.1)**	.15 **(60.0)**
	MSI-QTR-10	.10 (58.8)	.24 **(66.7)**	.19 (52.8)	.15 **(60.0)**
Average	QTR	.10 (58.8)	.21 (58.3)	.20 (55.6)	.15 (60.0)
	MSI-QTR-1	.13 (76.5)	.26 (72.2)	.26 (72.2)	.18 (72.0)
	MSI-QTR-5	.14 **(82.4)**	.29 **(80.6)**	.27 **(75.0)**	.19 **(76.0)**
	MSI-QTR-10	.14 **(82.4)**	.29 **(80.6)**	.26 (72.2)	.19 **(76.0)**

Table 9.7: Precision/Recall for the OHSUMED Collection Using Query Translation Based on Morpho-Semantic Normalization (% of Baseline in Brackets, Best Results Marked Bold). MSI-QTR-n Corresponds to MSI-QTR with n Disjunctive Queries.

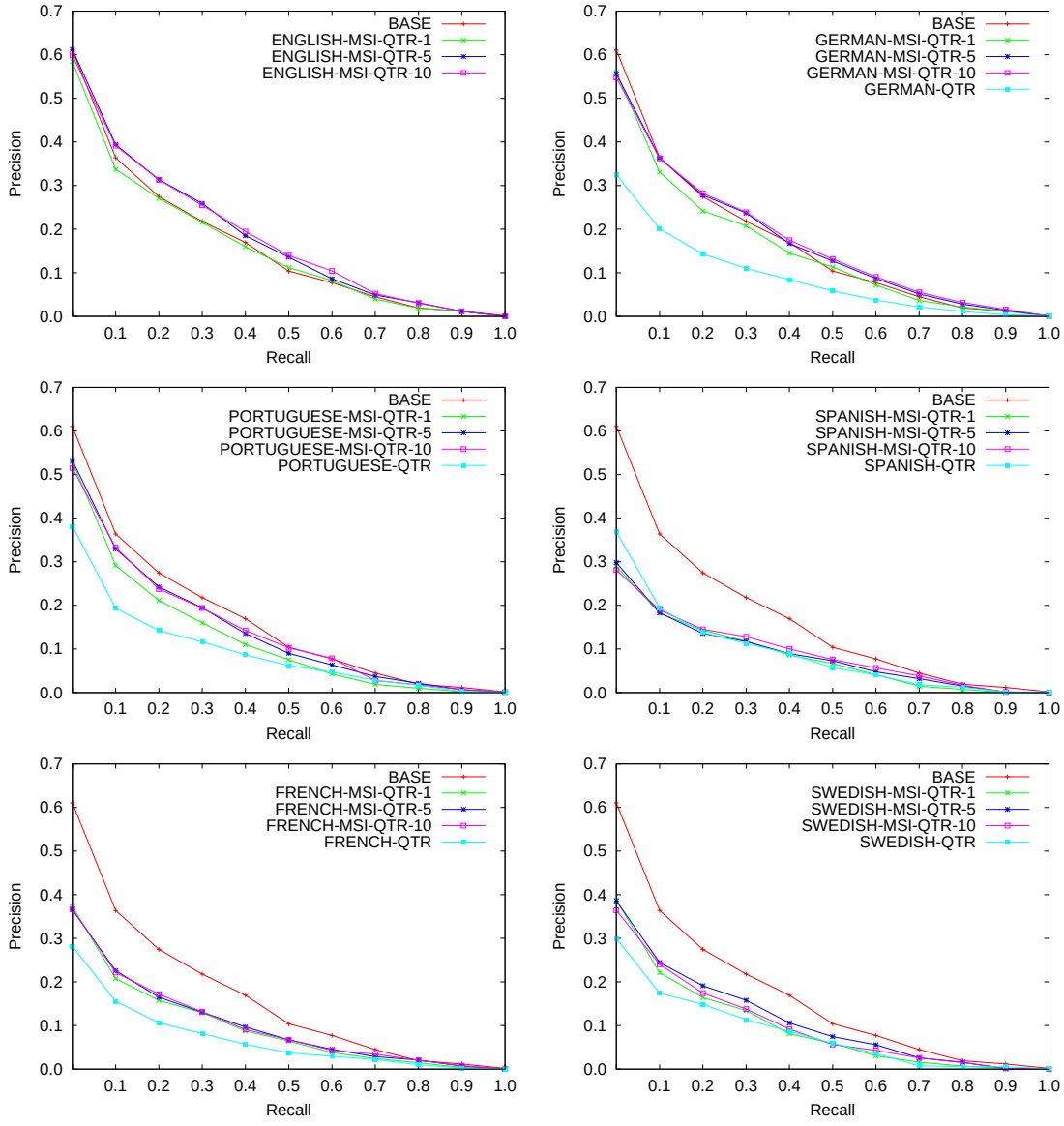

Figure 9.6: Precision/Recall Graphs for the OHSUMED Collection Using Query Translation Based on Morpho-Semantic Normalization

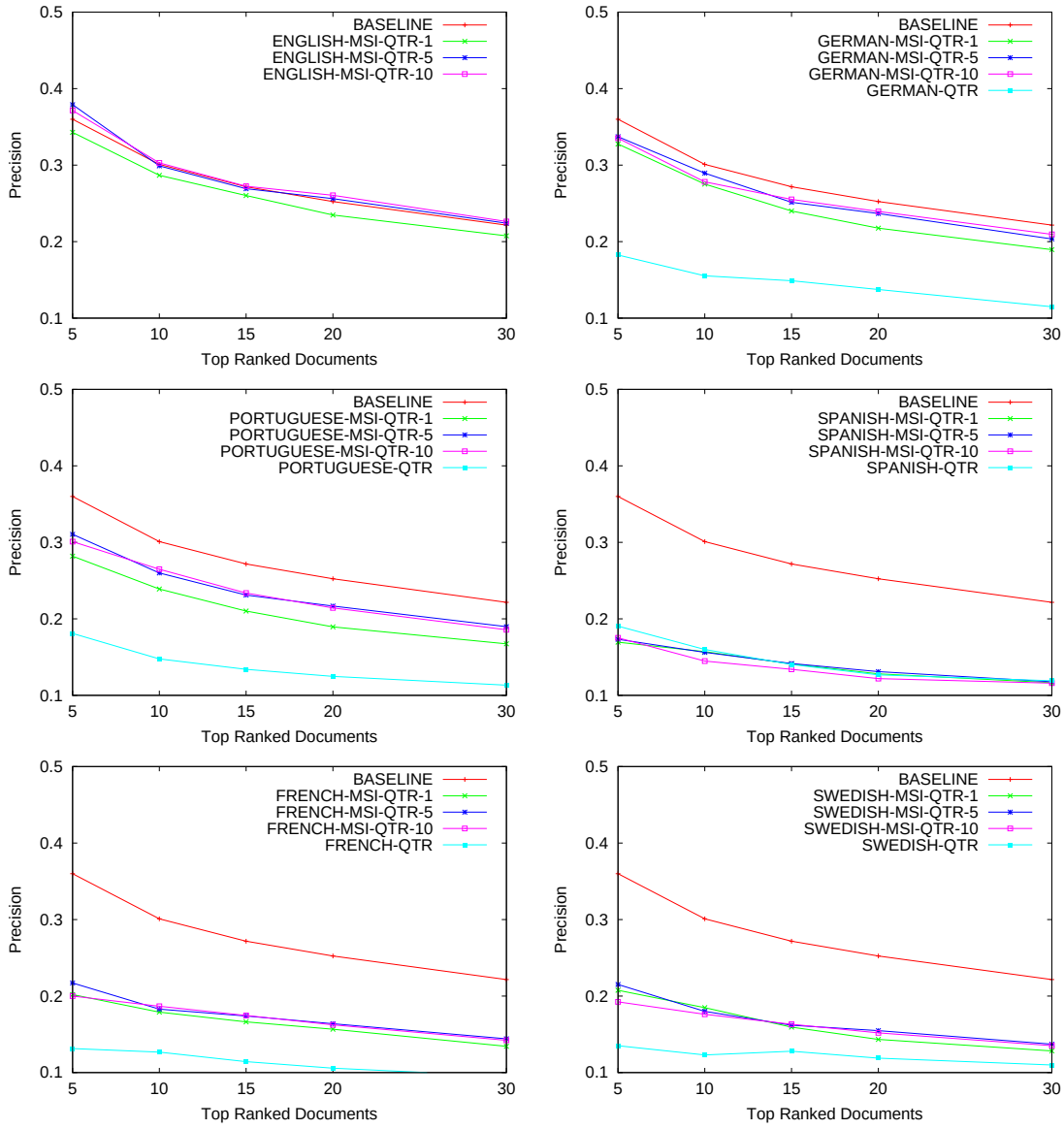

Figure 9.7: Exact Precision Graphs for the OHSUMED Collection Using Query Translation Based on Morpho-Semantic Normalization

For MSI-QTR-5, where five disjunctive queries are submitted to the search engine (cf. Table 9.6), precision increases for all languages except for Spanish, where no improvements are observable. The baseline is exceeded by 12% in the English scenario and for German, 100% of the monolingual baseline is reached. Extending the query by a total of ten disjunctions, precision can further be increased for German (106%) and French (65%).

Except for Spanish, where all MSI-QTR conditions yield the same result as QTR (53%), the query translation based on MorphoSaurus clearly outperforms QTR. In the German scenario, precision is even doubled (53% vs. 106%).

Figure 9.6 visualizes the data for all eleven standard recall points, while Figure 9.7 shows the exact precision scores for a few top ranked documents. To summarize, while QTR reaches an average relative precision of 59%, MSI-QTR-1 reaches 77%. Extending the queries by a total of five or ten disjuncts, precision reaches 82% with respect to 11pt average. While MSI-QTR-5 and MSI-QTR-10 also yield the same precision values for the average at the top two recall points and for top 20, allowing five disjunctive queries performs best considering the exact precision scores for the top 5 ranked documents. This is, at least partly, inline with current (controversial) research findings on query expansion (Gey & Chen, 2000).

9.5 ImageCLEFMed Results

Table 9.8 and Figures 9.8 and 9.9 show the corresponding results for the cross-validation using the ImageCLEFMed document collection.

Here, except for French, query expansion using a total of five or ten disjunctive queries outperforms all other conditions (11pt average). While standard query translation (QTR) reaches average scores of 0.06 (38%) for Spanish up to 0.13 (81%) for Portuguese and Spanish, MSI-QTR-10 even yields 100% of the baseline for Portuguese and 113% for German. This is particular surprising since in the latter case German outperforms English in the MSI-QTR-10 scenario (100%). For French, MSI-QTR values exceed QTR only considering the top 5 ranked documents. In other cases, QTR reaches higher or equal scores, compared to French MSI-QTR scenarios.

Language	Condition	11pt	top 2 pt	top 5	top 20
English	BASE	.16	.33	.39	.30
English	MSI-QTR-1	.14 (87.5)	.28 (84.8)	.39 (100)	.29 (96.7)
	MSI-QTR-5	.15 (93.8)	.34 (103.0)	.43 **(110.3)**	.34 **(113.3)**
	MSI-QTR-10	.16 **(100.0)**	.36 **(109.1)**	.43 **(110.3)**	.32 (106.7)
German	QTR	.09 (56.3)	.22 (66.7)	.27 (69.2)	.19 (63.3)
	MSI-QTR-1	.15 (93.8)	.29 (87.9)	.39 (100.0)	.25 (83.3)
	MSI-QTR-5	.17 (106.3)	.33 (100.0)	.43 (110.3)	.28 (93.3)
	MSI-QTR-10	.18 **(112.5)**	.36 **(109.1)**	.45 **(115.4)**	.32 **(106.7)**
Portuguese	QTR	.13 (81.3)	.20 (60.6)	.31 (79.5)	.21 (70.0)
	MSI-QTR-1	.15 (93.8)	.29 (87.9)	.35 (89.7)	.27 (90.0)
	MSI-QTR-5	.16 **(100.0)**	.32 (97.0)	.39 (100.0)	.28 (93.3)
	MSI-QTR-10	.16 **(100.0)**	.35 **(106.1)**	.43 **(110.3)**	.29 **(96.7)**
Spanish	QTR	.13 (81.3)	.24 (72.7)	.29 (74.4)	.21 (70.0)
	MSI-QTR-1	.13 (81.3)	.27 (81.8)	.30 (76.9)	.22 (73.3)
	MSI-QTR-5	.14 **(87.5)**	.30 **(90.9)**	.34 (87.2)	.24 (80.0)
	MSI-QTR-10	.13 (81.3)	.30 **(90.9)**	.36 **(92.3)**	.25 **(83.3)**
French	QTR	.09 **(56.3)**	.16 **(48.5)**	.23 (59.0)	.16 **(53.3)**
	MSI-QTR-1	.07 (43.8)	.16 **(48.5)**	.21 (53.8)	.15 (50.0)
	MSI-QTR-5	.07 (43.8)	.15 (45.5)	.25 **(64.1)**	.16 **(53.3)**
	MSI-QTR-10	.07 (43.8)	.16 **(48.5)**	.20 (51.3)	.16 **(53.3)**
Swedish	QTR	.06 (37.5)	.13 (39.4)	.19 (48.7)	.13 (43.3)
	MSI-QTR-1	.12 (75.0)	.23 (69.7)	.28 (71.8)	.21 (70.0)
	MSI-QTR-5	.13 **(81.3)**	.27 (81.8)	.39 **(100.0)**	.25 **(83.3)**
	MSI-QTR-10	.13 **(81.3)**	.30 **(90.9)**	.38 (97.4)	.25 **(83.3)**
Average	QTR	.11 (68.8)	.21 (63.6)	.28 (71.8)	.20 (66.7)
	MSI-QTR-1	.13 (81.3)	.25 (75.8)	.32 (82.1)	.23 (76.7)
	MSI-QTR-5	.14 **(87.5)**	.29 (87.9)	.37 (94.9)	.26 (86.7)
	MSI-QTR-10	.14 **(87.5)**	.30 **(90.9)**	.38 **(97.4)**	.27 **(90.0)**

Table 9.8: Precision/Recall for the ImageCLEFMed Collection Using Query Translation Based on Morpho-Semantic Normalization (% of Baseline in Brackets, Best Results Marked Bold). MSI-QTR-n Corresponds to MSI-QTR with n Disjunctive Queries.

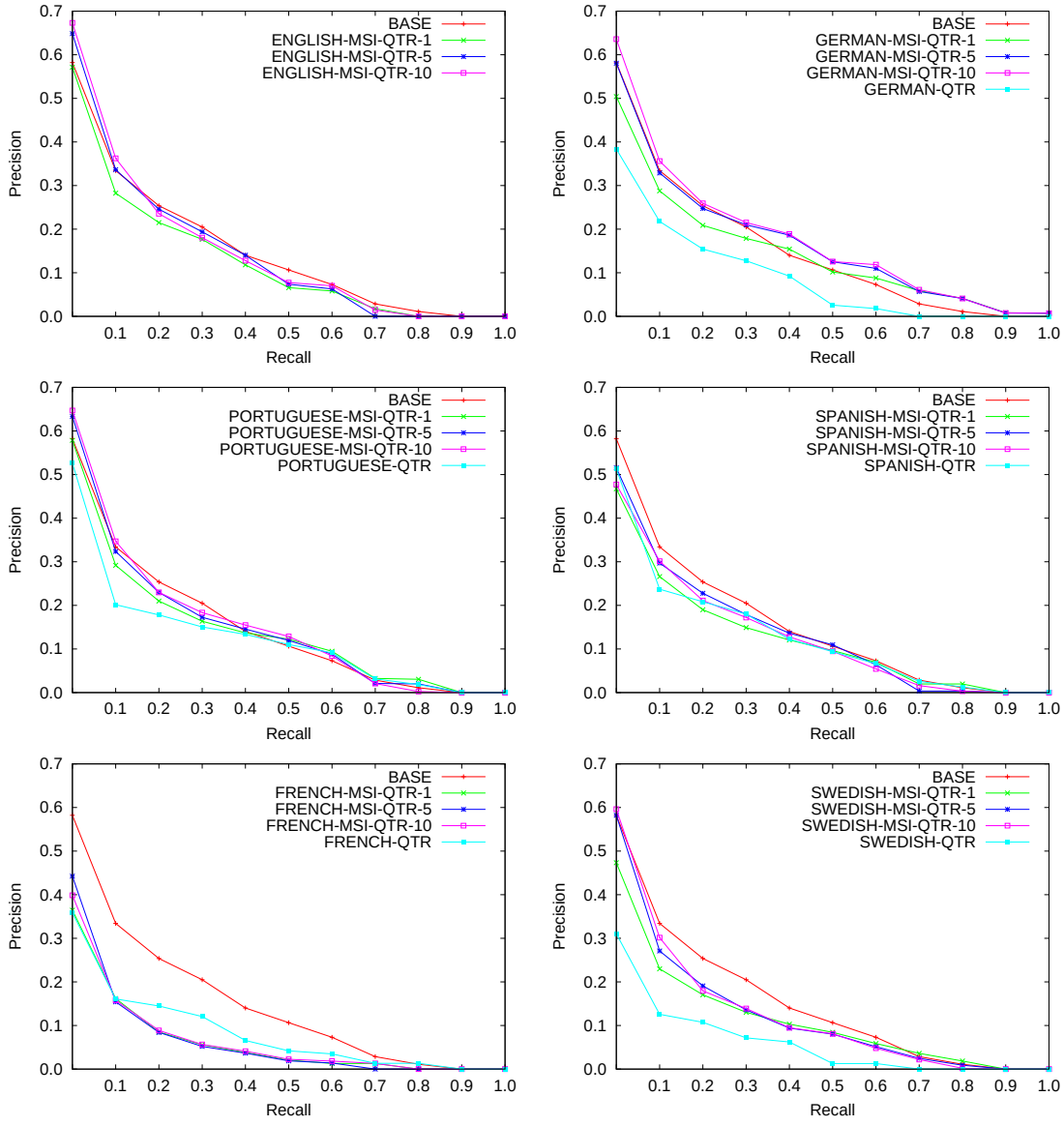

Figure 9.8: Precision/Recall graphs for the ImageCLEFMed Collection Using Query Translation Based on Morpho-Semantic Normalization

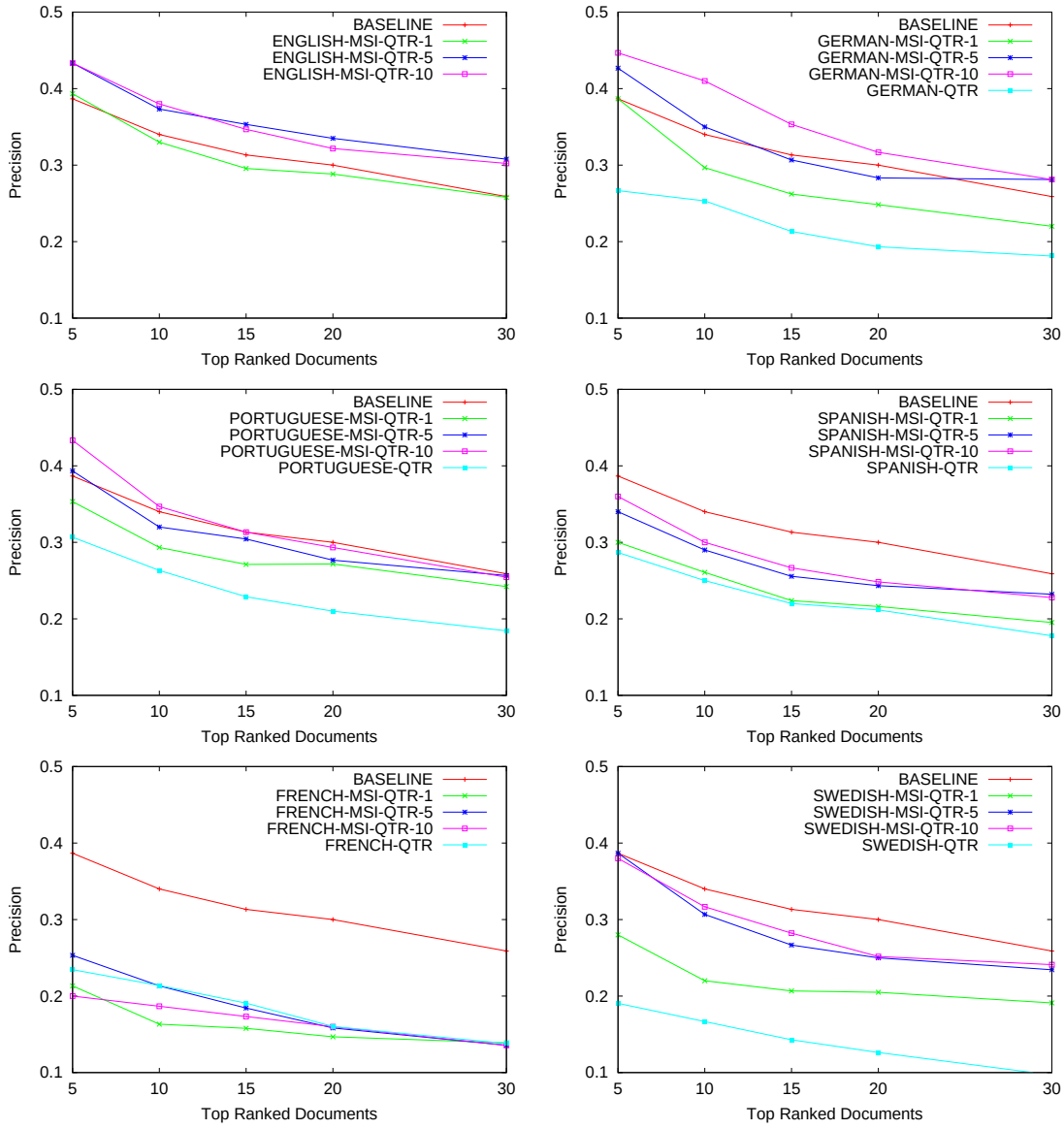

Figure 9.9: Exact Precision graphs for the ImageCLEFMed Collection Using Query Translation Based on Morpho-Semantic Normalization

Averaged over all languages, QTR has a mean average precision score of 0.11 (69%), while MSI-QTR-1 yields 0.13 (81%). Using query expansion (five or ten disjuncts) additionally increases the performance to 88% of the baseline. Considering only a few top ranked documents, MSI-QTR-10 achieves best results, with a gain of 3.3 percentage points over MSI-QTR-5 and a boost of 13 percentage points over QTR-MSI-1 while standard query translation (QTR) performs 23 percentage points worse.

9.6 Discussion

A comprehensive architecture for query translation is provided by Levow et al. (2005). They demonstrate the impact of various CLIR techniques using large-scale test collections in several languages (English → {French, Arabic, Chinese, German}). and found significant improvements in retrieval effectiveness for German and Chinese if subword segmentation is used (a gain of 35% for German and 14% in Chinese).

The use of n-gram techniques was reported as a successful approach for CLIR in the work of Savoy (2003a) and McNamee & Mayfield (2004). They demonstrate how overlapping character n-gram tokenization can provide retrieval accuracy that rivals the best current language-specific approaches for European languages.

Kamps et al. (2003) contrasted the effectiveness of language dependent approaches to document retrieval with language-independent approaches for nine European languages. They showed that morphological normalization improves retrieval effectiveness especially for languages that have a more complex morphology than English and that n-gram-based retrieval can be a viable option in the absence of linguistic resources.

A specific n-gram technique called *targeted s-gram* is analyzed by Pirkola et al. (2002) for English, German, and Swedish queries that were matched against their Finnish variants. They showed that their approach outperformed the conventional n-gram matching techniques particularly for short words and short longest common subsequences.

In contrast to these approaches which are based on the processing of *character* n-grams, *subword* n-grams are used here for performing the language transfer for CLIR. It has been shown that such a sophisticated approach outperforms a simple machine translation approach at large.

Compared to the evaluation results of the approach in which both queries and documents are transformed into the MorphoSaurus interlingua (Chapter 8) the outcome of query translation using morpho-semantic indexing as proposed here is slightly inferior. Still, one has to consider the fact that the interlingual transformation of huge and variable document collections (or the Web, in general) is realistically speaking not manageable. Thus, the approach including both query translation and a subsequent connection to a standard internet search engine offers a very good alternative to the MorphoSaurus core technology when huge external document collections have to be considered.

Chapter 10

Multilingual MeSH Mapping

Manual indexing or categorization of documents requires skillful human experts to perform a routine task, *viz.* to assign index terms or classification codes (usually, taken from a controlled vocabulary) to documents (journal or newspaper articles, technical reports, etc.). Constraining the choice of allowed descriptors to those organized in a thesaurus (e.g., the Medical Subject Headings (MeSH, 2005) or the Unified Medical Language System (UMLS, 2005)) creates additional benefits in so far that the document space is structured by semantically related areas. As a consequence, search capabilities become more powerful, e.g., by query expansion that incorporates synonyms, semantically more specific terms, etc. Large bibliographic services such as the retrieval system PubMed (the online interface to MEDLINE and related databases) mainly rely on the intellectual indexers' performance as far as the content description of documents is concerned.

The manual assignment of index terms out of a very large set of descriptors is not only a laborious and often tedious task but also one that is quite expensive. This is also evidenced by the NLM which, in the nineties, spent over two million dollars and employed 44 full-time equivalent indexers each year on that task (Hersh et al., 1994b).

MEDLINE covers English as well as non-English documents, though the indexing is in English only. Up until now, more than 14 million bibliographic units have been indexed and classified using the English version of the MeSH as a controlled

vocabulary. A few terminology mappers exist from English to some non-English languages, but their coverage is far from being complete (given the English MeSH). Because the physicians' native languages are much more dominant than in other scientific disciplines, the focus on English as medical content description language creates a serious bottleneck for tentative users of PubMed in non-English-speaking countries.

In order to reuse this bulk of intellectual work for languages other than English, MorphoSaurus can be used to learn from that data-rich experience in the following way: Assuming that the English indexing of medical documents is a highly esteemed asset, lexical patterns are determined from the abstracts and related to their associated index terms. Once lexical items can be mapped from the languages covered by MorphoSaurus to their English lexical correlates, English indexing patterns (together with the non-English ones) can be reused for the non-English language in focus given the mediating interlingua. Hence, the methodology proposed here learns from the past (English) indexing experience and transfers it in an unsupervised way to non-English languages, as well (Markó et al., 2003; Markó et al., 2004a; Hahn et al., 2004b).

10.1 Learning Indexing Patterns

In the following, a statistical, a heuristic and a hybrid approach is described to automatically assign English MeSH entries as document descriptors for English as well as German, Portuguese, Spanish and French MEDLINE documents given sets of *a priori* assigned index terms to English documents.[1] MeSH consists of sets of terms denoting descriptors in a hierarchical structure. In its 2006 version, nearly 24,000 so-called main headings with over 145,000 synonyms (entry terms) occur. Figure 10.1 (bottom) shows the tree structures for the MeSH main headings *Femoral Neck Fractures* and *Sepsis* which, amongst others, have been manually assigned to a MEDLINE abstract (taken from PubMed, cf. Figure 10.1, top). While the first mapping from title words to a MeSH term simply requires the consideration of sin-

[1] Unfortunately, to the best of knowledge, there are no Swedish abstracts linked to MEDLINE.

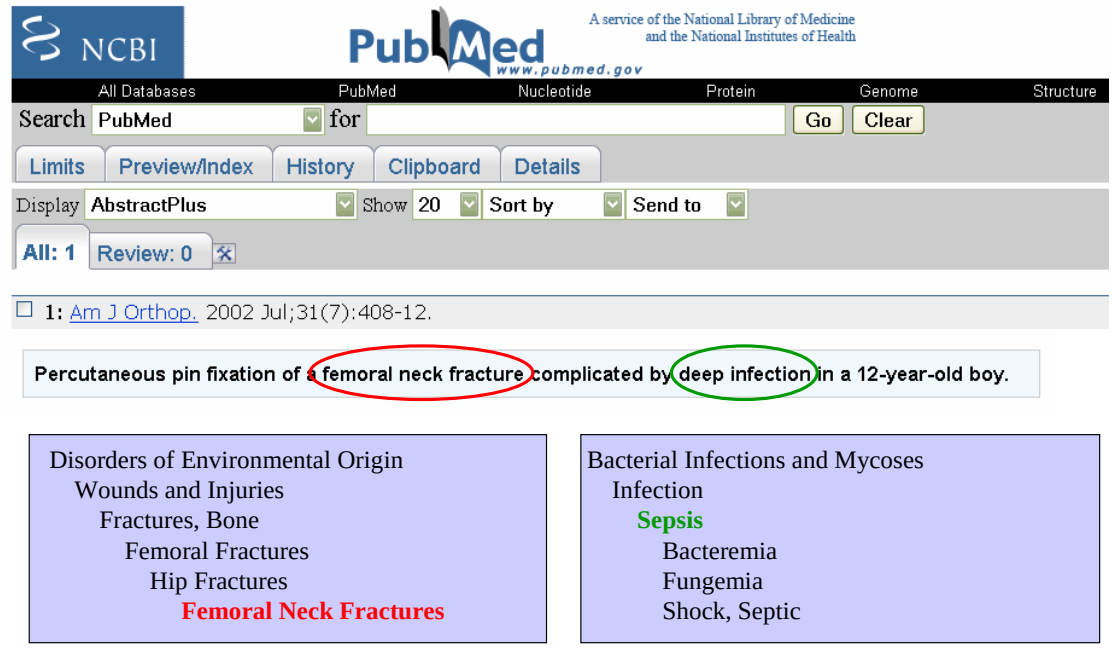

Figure 10.1: Sample Assignment of MeSH Descriptors to MEDLINE Abstracts

gular/plural forms, the second one is not that straightforward: Here, the (human) indexer decided to map *deep infection* to the MeSH heading *Sepsis*, instead of its parent node *Infection*. This association is not only motivated by the world knowledge of human curators, but also by the fact that the indexers usually have access to the whole publication, not only to titles and abstracts. Nevertheless, such kind of associations can be identified by well-known statistical machine learning methods. The particularity of the MorphoSaurus approach to assign descriptors to documents is based on using such methods together with the interlingual representation for the documents as well as the indexing vocabulary. Using this representation has the advantage of training the indexing system on texts written in one or more language(s) and testing them with documents of any (other) language. This method allows the processing of documents in any language covered by MorphoSaurus lexicons.

Language	Articles	Words
English	35,000	7,291,239
German	3,508	576,463
Portuguese	862	154,866
Spanish	998	250,268
French	8,025	1,591,578
Total	48,493	9,864,414

Table 10.1: Training Corpus Statistics for Statistical MeSH Mapping

10.1.1 Statistical MeSH Mapping

The starting point of the method to statistically assign MeSH terms to documents is to collect medical abstracts from MEDLINE, to which English MeSH main headings have already been manually assigned. For English, a subset of 35,000 documents was taken (Table 10.1). The training material for the other languages, i.e. (non-English) articles which are linked to MEDLINE and assigned with (English) MeSH terms, are much smaller due to limited availability (also cf. Section 5.1.1.1).

The algorithm then processes the sample of medical abstracts ($word_1 \ldots word_m$ in Figure 10.2, step A) taken from MEDLINE, to which English MeSH main headings have already been assigned manually ($MeSH_x$ and $MeSH_y$ in Figure 10.2, step A). Documents are morpho-semantically normalized, thus transforming the original document into a sequence of MIDs ($MID_1 \ldots MID_n$ in Figure 10.2, step B). Based on that representation a Bayesian approach is pursued which ignores the *a priori* probabilities of the descriptors. Thus, statistical evidence for class identifier (MID) trigrams is computed by basically counting their frequency of co-occurrence in the training corpus with individual (manually supplied) MeSH entries (Figure 10.2, step C). This is how indexing patterns are 'learned'.

In the test phase, when aiming at extracting MeSH terms as valid descriptors for a normalized document (cf. Figure 10.2, step D), these terms are ranked by their

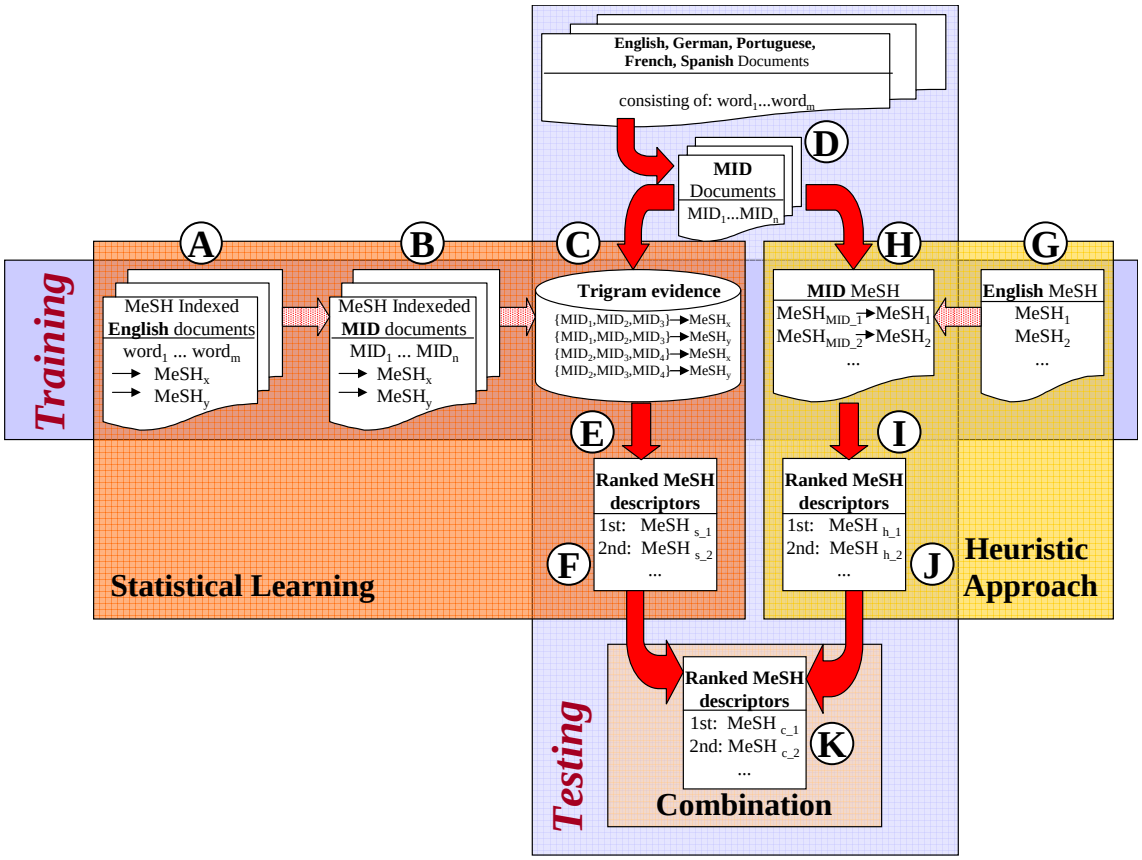

Figure 10.2: Architecture of the Combined Indexing System

weighting (w) values calculated as follows:

$$w(MeSH_i \mid MID_1, ..., MID_n) = log \prod_{j=1}^{n-2} \begin{cases} \frac{P(MID_j, MID_{j+1}, MID_{j+2} | MeSH_i)}{P(MID_j, MID_{j+1}, MID_{j+2})} & \text{, if defined} \\ 1 & \text{, otherwise} \end{cases}$$

Given a document which contains n class identifiers (MIDs), the conditional weighting value for $MeSH_i$, a particular MeSH main heading, is computed by the product of the computed conditional probabilities P of the MID trigrams in the text that co-occur with the descriptor $MeSH_i$ in the training set, divided by the *a priori* probability of the corresponding text trigrams in the training collection, if both probabilities can be observed, at all (cf. Figure 10.2, step E). Here, the denominator takes into account the fact that infrequent terms have a greater explanatory power for a given entity when faced with large quantities of data and, hence, increase

the weighting value for that entity. If no trigram that is currently being processed appears in the training data, or if it is not associated with the current descriptor $MeSH_i$, it remains neutral (multiplication with 1). This expresses the fact that there is no evidence for a further refinement (simply because a combination of a trigram and a MeSH descriptor missing in the training set does not mean that it may never occur, at all).

MID trigrams are treated in an unordered way. They are defined as a set of MIDs that co-occur within a document window of three text tokens, regardless of the original sequence of words that produced the set of MIDs. The reason for this is that in many languages the MID order changes when genitives or prepositions come into play, as with *"femoral neck fracture"* vs. *"fractured neck of femur"* corresponding to $[\#femur, \#neck, \#fractur]$ vs. $[\#fractur, \#neck, \#femur]$. Finally, all extracted MeSH descriptors, $MeSH_{s_1}$, $MeSH_{s_2}$ ..., are ranked according their weighting value (Figure 10.2, step F).

10.1.2 Heuristic MeSH Mapping

The heuristic approach only relies on the MeSH Thesaurus and a collection of documents. Based on a set of heuristic criteria, a fully automatic MeSH indexing of the documents is computed. Unlike the learning method, no prior indexing of documents is necessary. In the training phase, all English MeSH main headings, $MeSH_1$, $MeSH_2$, etc., (cf. Figure 10.2, step G) undergo the morpho-semantic normalization procedure. Hence, all words covered by the English lexicon are substituted by their corresponding unique MIDs resulting in the morpho-semantically normalized representations $MeSH_{MID\_1}$, $MeSH_{MID\_2}$, etc., which are linked to the original MeSH descriptors (cf. Figure 10.2, step H).

In the test phase, English, German, Portuguese, Spanish and French documents, defined by a sequence $word_1 \ldots word_m$, are processed by the morphological engine resulting in a sequence $MID_1 \ldots MID_n$ at the interlingua layer (Figure 10.2, step D). Afterwards, heuristic rules (some of them already proposed by NLM's indexing initiative (Aronson et al., 2000)) are applied to the normalized test documents. In essence, this means that each MeSH descriptor whose normalized representation

contains at least one of the MIDs in the document is retrieved. Next, each normalized MeSH descriptor is assessed against the normalized text by computing diverse factors (Figure 10.2, step I). The most important metrics are:

- **Longest Match Factor:** On the level of MIDs, individual MeSH descriptors, which appear as single entries, can also appear together in additional MeSH entries. For example, the German term *"Bauchschmerzen"* (*"abdominal pain"*) that appears in a text and is normalized to the MIDs *#abdom* and *#pain* is, amongst others, associated to the MeSH entries "Abdominal Pain" ([*#abdom*, *#pain*]), "Abdomen" (*#abdom*) and "Pain" (*#pain*). If two or more normalized MeSH descriptors can be merged to one longer MeSH descriptor, the latter is preferred over its constituents.

- **Phrase Factor:** The number of different MIDs in a phrase that match the MIDs in a normalized descriptor is called *MID count*. In addition, the phrase interval of a normalized descriptor can be considered as the span between the first and the last MID associated with this descriptor in a phrase. The *phrase factor*, then, is defined as the ratio of MID count and phrase interval. So, the Portuguese phrase *"o fígado do paciente foi transplantado"* (*"the patient's liver was transplanted"*) will be transformed into [*#hepat*, *#patient*, *#transplant*]. Given the normalized descriptor for "liver transplantation" ([*#hepat*, *#transplant*]), the corresponding MID count is 2, the phrase interval amounts to 3. So, the phrase factor equals 2/3.

- **Entry Factor:** The *entry factor* is the MID count divided by the number of MIDs of the associated descriptor. For example, the German noun phrase *"noduläre Hyperplasie"* (*"nodular hyperplasia"*) is normalized to [*#nodul*, *#above*, *#plast*] and the MeSH descriptor "Focal Nodular Hyperplasia" to [*#focal*, *#nodul*, *#above*, *#plast*]. The corresponding entry factor is 3/4.

- **Title Factor:** A descriptor found in the title will be ranked higher than others.

Finally, all possible descriptors are ordered according to a weighted average of the above metrics ($MeSH_{h\_1}$, $MeSH_{h\_2}$,... in Figure 10.2, step J).

10.1.3 Hybrid Approach

The statistical learning of indexing patterns and the heuristic add-ons were pooled in order to find out whether a combined effort performs better than any of the two in isolation. Hence, both approaches were merged in the following way. First, all descriptors that are ranked in the top 30 by both of the methods are set to the top of the result list ($MeSH_{c\_1}$, $MeSH_{c\_2}$... in Figure 10.2, step K). After the first k positions (30≥k) have been populated that way, the remaining positions are incrementally filled by the following rule: Two entries on the top of the output of the statistical approach are alternately incorporated into the final result, followed by one entry of the heuristic approach, until both lists (maximum length: 100 terms) are exhausted. Previous experiments have shown that this empirically motivated procedure leads to much more favorable results than a formal one, e.g., by multiplying the outcome values of the different weighting functions.

10.2 Evaluation

Text collections were randomly assembled for the training phase for the statistical learning of indexing patterns and the test phase (500 abstracts for each language considered, cf. Table 10.2). The data acquired from the training phase were then used for the indexing of English, German, Portuguese, Spanish and French documents. The indexing results were evaluated against the manually supplied MeSH main headings. This data serves as the *de facto* gold standard for the experiments (similar to the study of the indexing initiative of the NLM (Aronson et al., 1999)). Unfortunately, the human indexing results in MEDLINE are not really consistent. Funk & Reid (1983) measured 54.5% interrater agreement with regard to manually assigned MeSH main headings for English abstracts (41.5% for German abstracts). Obviously, such inconsistencies in the test collection will also affect the validity of the evaluation results when taking this data as gold standard.

Language	Articles	Words
English	500	103,681
German	500	80,684
Portuguese	500	88,674
Spanish	500	122,880
French	500	96,310
Total	2,500	492,229

Table 10.2: Test Corpus Statistics for Statistical MeSH Mapping

In earlier experiments the performance of the three different methods, viz. heuristic mapping, statistical mapping, and the combination of both were compared considering only a smaller set of German abstracts covering clinical disciplines only, both for training as well as testing (Markó et al., 2003). In a subsequent study this evaluation was repeated on different subsets of MEDLINE covering the whole MeSH using only English documents for training and English/German/Portuguese documents for testing (Markó et al., 2004a). Now, the effect of combining monolingual training data to multilingual evidence via the interlingua is analyzed more detailed.

In particular, the following experimental conditions are considered:

- **H core:** the heuristic approach to MeSH mapping using the core engine of MorphoSaurus
- **S core:** the statistical approach using the core engine of MorphoSaurus and monolingual training data
- **S D:** the statistical approach using MorphoSaurus with disambiguation module and monolingual training data
- **S full:** the statistical approach using MorphoSaurus with disambiguation and acronym resolution module and monolingual training data

- **S core +:** the statistical approach using the core engine of MorphoSaurus and multilingual training data
- **S D +:** the statistical approach using MorphoSaurus with disambiguation module and multilingual training data
- **S full +:** the statistical approach using MorphoSaurus with disambiguation and acronym resolution module and multilingual training data
- **M core +:** the mixed-mode, hybrid approach using the core engine of MorphoSaurus and multilingual training data
- **M D +:** the mixed-mode, hybrid approach using MorphoSaurus with disambiguation module and multilingual training data
- **M full +:** the mixed-mode, hybrid approach using MorphoSaurus with disambiguation and acronym resolution module and multilingual training data

10.3 Results

Table 10.3 (English, German), Table 10.4 (Portuguese, Spanish) and Table 10.5 (Swedish and average values) depict the precision and recall values for the chosen test scenarios, for which the top 5, 10 and 50 ranked descriptors are considered.

When examining average values only (Table 10.5, bottom) and focusing on the top five proposed descriptors, between 9% (S core) and 17% (M D +) of all relevant MeSH terms are retrieved at a precision rate of 20% (S core) to 40% (M D +). Looking at the top 50 of the system-generated descriptors, precision drops to between 5% (heuristic approach) and 13% (M D +), while recall increases to between 22% (heuristics only) and 56% (M D +). For the top 5 proposed descriptors, the heuristic approach (23% precision at 10% recall) performs slightly better than the statistical approach using monolingual training data only (20% precision at 9% recall for S core). However, when considering the MeSH terms in the top 10 or top 50, even the simple statistical approach outperforms the heuristic one. By applying the disambiguation module, precision (recall) can further be increased by up to

Language	**Method**	**Top 5**		**Top 10**		**Top 50**	
		Prec	Rec	Prec	Rec	Prec	Rec
English	H core	0.34	0.13	0.23	0.17	0.07	0.28
	S core	0.33	0.13	0.27	0.20	0.12	0.44
	S D	0.37	0.14	0.29	0.22	0.12	0.45
	S full	0.32	0.12	0.25	0.19	0.11	0.40
	S core +	0.34	0.13	0.27	0.21	0.12	0.44
	S D +	0.38	0.14	0.29	0.22	0.12	0.47
	S full +	0.34	0.13	0.26	0.20	0.11	0.42
	M core +	0.43	0.16	0.34	0.26	0.14	0.52
	M D +	**0.45**	**0.17**	**0.36**	**0.27**	**0.15**	**0.55**
	M full +	0.43	0.16	0.34	0.26	0.14	0.52
German	H core	0.27	0.11	0.18	0.15	0.06	0.25
	S core	0.20	0.09	0.17	0.14	0.09	0.37
	S D	0.25	0.10	0.20	0.17	0.10	0.41
	S full	0.23	0.10	0.19	0.16	0.09	0.39
	S core +	0.32	0.14	0.26	0.22	0.12	0.48
	S D +	0.38	0.16	0.31	0.26	0.13	0.55
	S full +	0.36	0.15	0.28	0.24	0.12	0.52
	M core +	0.38	0.16	0.29	0.25	0.13	0.53
	M D +	**0.41**	**0.17**	**0.33**	**0.28**	**0.14**	**0.59**
	M full +	0.38	0.16	0.29	0.25	0.13	0.53

Table 10.3: Precision/Recall Table for English and German Using Different Indexing Methods at Different Cut-off Points (Best Results Marked Bold)

Language	**Method**	**Top 5**		**Top 10**		**Top 50**	
		Prec	Rec	Prec	Rec	Prec	Rec
Portuguese	H core	0.22	0.10	0.15	0.14	0.05	0.23
	S core	0.10	0.05	0.10	0.09	0.08	0.37
	S D	0.12	0.06	0.11	0.11	0.08	0.37
	S full	0.13	0.06	0.12	0.11	0.08	0.37
	S core +	0.26	0.12	0.21	0.20	0.10	0.46
	S D +	0.32	0.15	0.27	0.25	0.11	0.53
	S full +	0.32	0.15	0.26	0.24	0.11	0.53
	M core +	0.32	0.15	0.25	0.23	0.11	0.50
	M D +	**0.36**	**0.17**	**0.29**	**0.27**	**0.13**	**0.59**
	M full +	0.32	0.15	0.25	0.23	0.11	0.50
Spanish	H core	0.11	0.05	0.08	0.07	0.03	0.12
	S core	0.21	0.09	0.17	0.15	0.08	0.33
	S D	0.25	0.11	0.21	0.18	0.08	0.34
	S full	0.25	0.11	0.21	0.18	0.08	0.33
	S core +	0.40	**0.18**	**0.30**	**0.26**	**0.12**	0.52
	S D +	**0.41**	**0.18**	**0.30**	**0.26**	**0.12**	0.53
	S full +	0.40	0.17	0.29	0.25	0.11	0.50
	M core +	**0.41**	**0.18**	0.29	0.25	**0.12**	0.53
	M D +	**0.41**	**0.18**	0.29	0.25	**0.12**	**0.54**
	M full +	**0.41**	**0.18**	0.29	0.25	**0.12**	0.53

Table 10.4: Precision/Recall Table for Portuguese and Spanish Using Different Indexing Methods at Different Cut-off Points (Best Results Marked Bold)

Language	**Method**	**Top 5**		**Top 10**		**Top 50**	
		Prec	Rec	Prec	Rec	Prec	Rec
French	H core	0.22	0.10	0.15	0.14	0.05	0.23
	S core	0.17	0.08	0.13	0.12	0.07	0.30
	S D	0.21	0.10	0.17	0.15	0.08	0.38
	S full	0.21	0.09	0.17	0.15	0.08	0.36
	S core +	0.24	0.11	0.19	0.17	0.09	0.43
	S D +	0.30	0.14	0.24	0.22	0.11	0.50
	S full +	0.28	0.13	0.22	0.21	0.10	0.48
	M core +	0.32	0.14	0.23	0.21	0.10	0.46
	M D +	**0.35**	**0.16**	**0.26**	**0.24**	**0.12**	**0.54**
	M full +	0.32	0.14	0.23	0.21	0.10	0.46
Average	H core	0.23	0.10	0.16	0.13	0.05	0.22
	S core	0.20	0.09	0.17	0.14	0.08	0.36
	S D	0.24	0.10	0.20	0.17	0.09	0.39
	S full	0.23	0.10	0.19	0.16	0.09	0.37
	S core +	0.31	0.13	0.25	0.21	0.11	0.47
	S D +	0.36	0.15	0.28	0.24	0.12	0.51
	S full +	0.34	0.15	0.26	0.23	0.11	0.49
	M core +	0.37	0.16	0.28	0.24	0.12	0.51
	M D +	**0.40**	**0.17**	**0.31**	**0.26**	**0.13**	**0.56**
	M full +	0.37	0.16	0.28	0.24	0.12	0.51

Table 10.5: Precision/Recall Table for Swedish and Average for all Languages Using Different Indexing Methods at Different Cut-off Points (Best Results Marked Bold)

4 (3, respectively) percentage points. When incorporating the acronym resolution module, which tends to add additional noise to the data (but enables cross-lingual comparisons), precision and recall decrease by at most two percentage points.

By pooling monolingual training data to multilingual evidence via the MorphoSaurus interlingua, considerable enhancements can be observed for the statistical approach (for the top 5, 12 percentage points precision gain for (S D) to (S D +) and 12 percentage points recall for the top 50). Especially for Portuguese, for which only few training material is available (cf. Table 10.1), additionally using English training data (that are easily to obtain) is a substantial benefit that can be expressed in terms of up to 20 percentage points performance increase (S D + compared to S D).

The different contributions of the two basic approaches (heuristic and statistical) were examined in more detail, as well. Only focusing on multilingual training data, the statistical learning approach always outperforms the heuristic one substantially, for all languages at all cut-off points with respect to recall and precision. The learned indexing patterns are, therefore, the driving force for the performance of the system. Pooling both approaches, however, yields additional, substantial, benefits. Performance values range from 37% precision at 16% recall (top 5, M core + and M full +) and 12% precision at 51% recall (top 50) up to 40% precision at 17% recall (top 5, M D +) and 13% precision at 56% recall (top 50).

Figure 10.3 summarizes the resulting precision values for the different languages for the top 5, 10, 20, 50 and top 100 proposed descriptors and visualizes the differences between the experimental conditions. Accordingly, Figure 10.4 shows the corresponding graphs for recall values.

With recall values ranging between 24% and 28% for each of the languages for the top ten assigned descriptors, results seem to be not so shiny. However, when comparing these values to the average agreement of human indexers (for English 5.45 descriptors, for German only 4.15 MeSH terms, according to Funk & Reid (1983)) the indexing system proposed here derives 2.75 less descriptors in average for English and only 1.35 less for German in a fully automatic indexing environment.

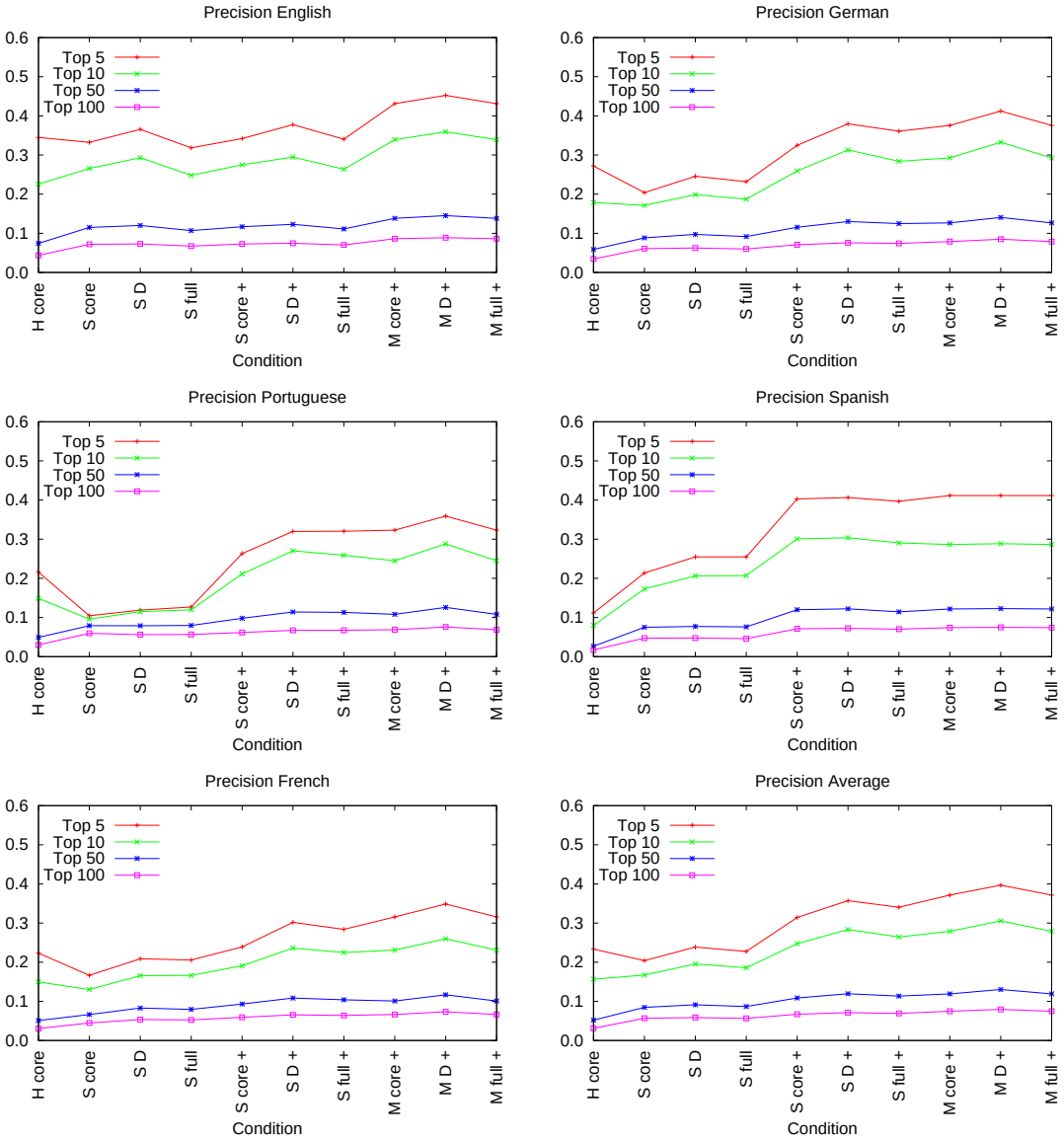

Figure 10.3: Exact Precision for MeSH Indexing

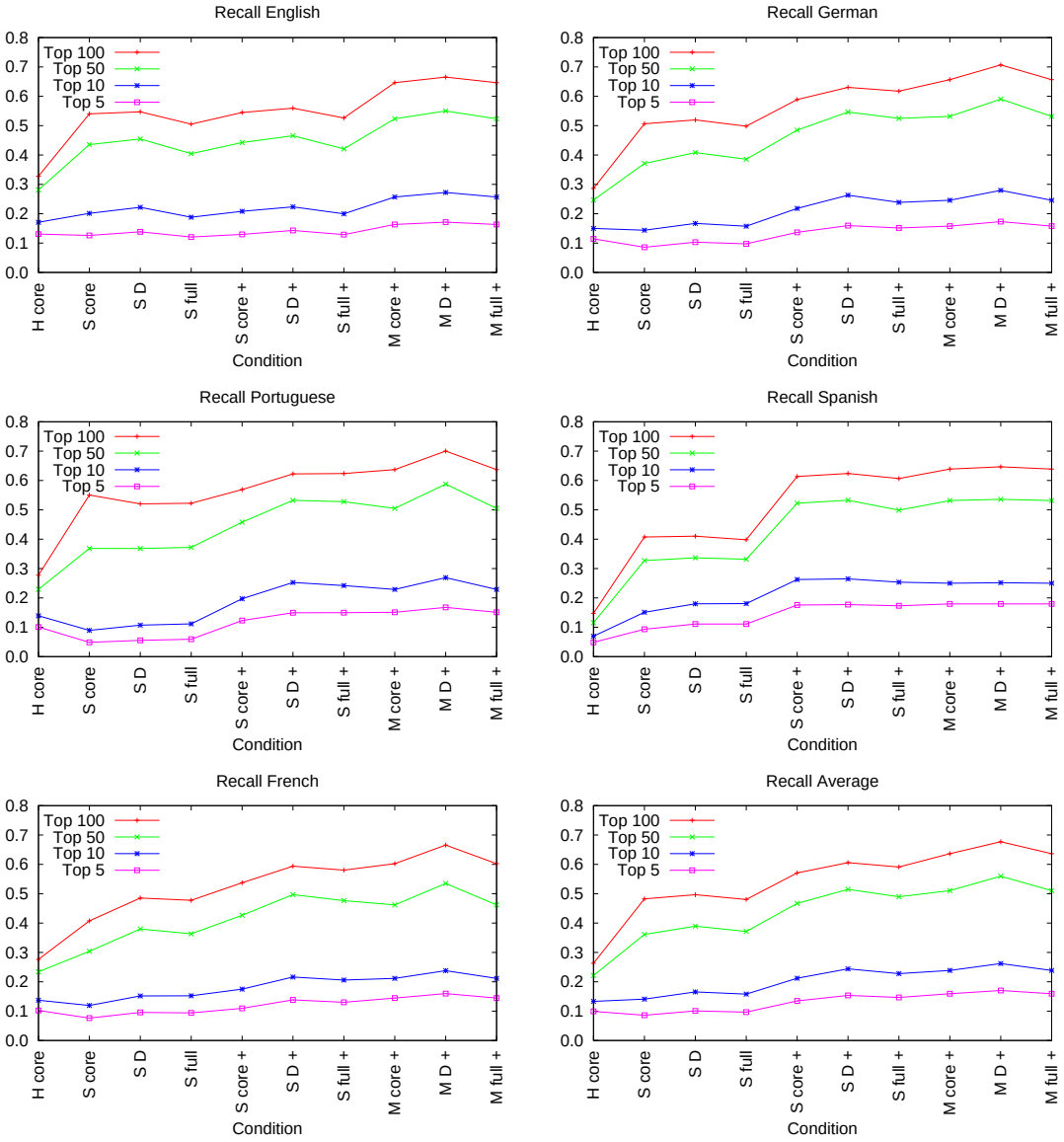

Figure 10.4: Exact Recall for MeSH Indexing

10.4 Discussion

The experiments of Hersh & Donohoe (1998) and Zweigenbaum et al. (2001) reveal the usefulness of incorporating morphological knowledge into automatic indexing procedures. At least for highly compounding languages such as German or Swedish, however, their proposed methods, *viz.* the enumeration of morphological variants in a semi-automatically generated lexicon (Zweigenbaum et al., 2001; Aronson, 2001) or the incorporation of a simple stemmer (Hersh & Donohoe, 1998) turn out to be inappropriate.

The (monolingual) MeSH mapping methods proposed by the NLM (Aronson et al., 2000; 1999) reach 48% precision for the top 10 proposed descriptors and 20% precision for the top 40 on a small test corpus comprised of 200 MEDLINE abstracts. 29% precision for the top 25 is reported after carrying out a re-evaluation on 273 MEDLINE articles (Aronson et al., 2004). As a comparison, using the approach proposed in this work, precision for English (M D +) varies between 36% (top 10) and 15% (top 50). By using full texts instead of abstracts only, the performance of an indexing system can be increased by 7%, as reported by Gay et al. (2005).

Névéol et al. (2005a) compared three different MeSH indexing systems for French by using 82 documents from CISMeF.[2] Both a regular expression-based approach and another one using TF/IDF measures retrieve 21% of all relevant MeSH terms with respect to the top 10 proposed descriptors. The third indexing system which is based upon the use of different medical terminological resources only achieves 13% precision. CISMeF also contains resources which are available in English and French, mostly coming from Canadian governmental websites such as Health Canada[3] and the Canadian Pediatric Society[4]. Using a subset of 51 documents, the French indexing system based on TF/IDF and the English one developed by the NLM have been evaluated in parallel. For English, 27% precision are reported for the top 10 results, while the performance of the French system is 23% (Névéol et al., 2005b).

[2]Catalogue and Index of Medical online resources in French, `http://www.cismef.org/`

[3]`http://www.hc-sc.gc.ca/`

[4]`http://www.cps.ca/`

Sebastiani (2002), however, emphasizes that performance comparisons of different evaluations have only limited value. Various experimental conditions have to be taken into account, *viz.* structure and size of the documents sets and the controlled vocabulary, choices of text preprocessing (e.g., morpho-semantic analysis *vs.* stemming), the indexing method being applied (e.g., rule-based *vs.* statistical), parameter tuning, etc. Nevertheless, he concludes that any content-based indexing method that incorporates some machine learning algorithm (e.g., probabilistic classifiers, decision tree classifiers) does better than methods without any learning device. Various experiments carried out on the REUTERS-21578 corpus, the most widely used benchmark collection in automated indexing (Rose et al., 2002), showed that the combination of different indexing methods seems to perform best, in general. These considerations are also backed up by the results in this work.

To the best of knowledge, no efforts on direct translingual document indexing have been made, up until now. Ferber (1997) and Pouliquen et al. (2003) both apply monolingual indexing techniques to various languages and use a multilingual controlled vocabulary, for which exact translations exist (the EUROVOC thesaurus and the OECD macrothesaurus (cf. Section 12.2), respectively). Their learning algorithms have to be adapted to each language-specific document collection. In contrast, the statistical approach proposed here 'learns' descriptor assignments mainly from an English corpus, for which training data are easily obtainable. Document descriptors can then be assigned to any text whose underlying language is covered by MORPHOSAURUS.

Chapter 11

Towards a General Multilingual Medical Lexicon

Lexicons, especially designed for natural language processing purposes, can generally be characterized along several dimensions. Firstly, lexicons can provide different amounts of lexical information, such as part of speech, number, gender and case. Secondly, the coverage of a lexicon, which often captures the terminology of a specialized domain, indicates for how many words of a (domain-specific) text collection lexical information is available. For translation dictionaries, finally, special attention is drawn to the multilingual dimension.

There is currently no large electronic dictionary in the medical domain which is characterized by a true multilingual dimension, relevant coverage, and substantial lexical information at the same time. Of course, with the UMLS Metathesaurus (UMLS, 2005) there already exists a widely used multilingual resource with high coverage in the medical domain. However, lexical information is missing for other languages than English.

For non-specialized domains, a remarkable effort for developing mono- and multilingual dictionaries has been made. For example, WordNet (Fellbaum, 1998) provides a good coverage for general English. It may be useful for covering lay terminology of medicine (Burgun & Bodenreider, 2001) or bio-medicine (Bodenreider et al., 2003), for example within a consumer-oriented health information system.

The European counterpart, EuroWordNet (Vossen, 1998) tends towards a multilingual system, but with considerably diverse levels of lexical coverage.

Whenever medical terminology has been addressed in the construction of a multilingual dictionary with substantial lexical information, it lacks reasonable coverage or has been developed as a demonstration prototype (Chiao & Zweigenbaum, 2002).

The MorphoSaurus subword lexicons, which align medical words in different languages on the subword-level, provide high coverage of medical terminology in different languages. But morpho-syntactic information such as part-of-speech, case, gender, etc. is completely missing in this resource. Nevertheless, morpho-semantic indexing can be used for linking different monolingual resources into a multilingual repository with high coverage (Markó et al., 2006a; 2006b).

Multilinguality means at least that corresponding entries in different languages are connected, which is a difficult task and raises simple questions and concerns open issues, like e.g., in which cases a translation relationship truly holds for lexical entities. Therefore, syntactic as well as semantic criteria have to be developed, or, at least, a consensus of different lexical input providers has to be found.

Of course, monolingual resources exist for different languages, so the first step to merge them is to create a common framework for the integration of lexical entities from different languages, with respect to their intrinsic peculiarities.

11.1 Interchanging Lexical Information

The Interchange Format is a convention about the way to exchange linguistic information entering in the building process of a medical multilingual lexicon (Baud et al., 2005). The basic idea is that the exchange of information is performed through the Interchange Format only, and each contributor of lexical resources is converting available data into that representation.

Table 11.1 lists the fields of the interchange format. The most important ones are the following:

- **Lng:** The language field determines to which language a particular entry belongs. Up until now, the values are: EN for English, FR for French, DE for German, LA for Latin, SV for Swedish, ES for Spanish and PT for Portuguese.

Field	Description	Definition
Lng	Language	the language to which pertains the present entry
Id	Multilingual Identifier	the unique identifier of this entry
Typ	Entry Type	one of the 4 allowed types of entry (B,C,S,T)
Err	Correctness	flag for correctness of this entry
Lem	Lemma	the entry in its basic form
Mul	Morpho-syntactic Features	the MULTEXT morpho-syntactic tag of the lemma
Frm	Inflected Form	any inflected form
Mfr	Features of Inflected Form	the MULTEXT morpho-syntactic tag of the inflected form
Inf	Inflection Model	language specific information
Mis	Language Specific Argument	to be used freely by provider of entries
Prt	Decomposition	the decomposition of a compound entry into its parts
Str	Head	the head word of the term
Ref	Reference Lemma	ID of its lemma's entry (if inflection form)
Exa	Typical Usage	a sentence presenting a typical usage of this entry
Com	Comment	any comment or warning about this entry

Table 11.1: Fields of the Lexicon Interchange Format

- **Id:** This argument specifies the unique identifier of the multilingual lexicon entry, made of the concatenation of the name of the input provider and a consecutive number.

- **Typ:** The type of entry defines either a *basic entry* (B), a *subword entry* (S), a *compound entry* (C) or a *term entry* (T). By definition, these types are mutually exclusive. The *basic entry* encodes single words of the language, generally without a space character in their lemma. The *subword entry* is a marker for parts of words entering in the composition of a *compound entry*. Therefore, a *subword entry* can generally not be used standalone and a *compound entry* is for words, which have been explicitly recognized as a composition of two or more *subword entries*. Finally, a *term entry* (T) describes a sequence of words, generally separated by the space character.

- **Lem:** The lemma is the representation of the entry in its basic form (singular, nominative for nouns; infinitive for verbs). It is supposed to be recoverable from any occurring form by an inflectional morphology process which is language dependent. There is exactly one unique basic form for any entry.

- **Mul:** The code for encoding morphological and syntactic information is defined as in the open standard MULTEXT.[1] Language dependent extensions of MULTEXT may be used.

- **Frm:** An entry that describes a specific inflected form that is linked to an entry for its lemma through the **Ref** field.

- **Mfr:** The morpho-syntactic features of the inflected form using MULTEXT exactly as for the **Mul** field.

- **Prt:** The decomposition of compound entries.

- **Ref:** If the entry consists of an inflected form, a unique ID of its lemma entry is given.

[1]Common Specifications and Notation for Lexicon Encoding and Preliminary Proposal for the Tagsets (`http://nl.ijs.si/ME/V3/msd/related/msd-multext/`)

11.2 Resources

After agreeing upon the Interchange Format, partners from five different institutions who are active in the medical domain, as well as in linguistics, collected their monolingual lexical resources. These are:

- the French UMLF lexicon from different French health-related organizations and the University Hospitals of Geneva, Switzerland (33,718 entries) (Zweigenbaum et al., 2005)
- an English medical lexicon from Linköping University, Sweden (22,686 entries) (Nyström et al., 2006)
- a Swedish medical lexicon from Linköping University (23,223 entries) (Nyström et al., 2006)
- a Swedish medical lexicon from Göteborg University, Sweden (6,786 entries)
- the German Specialist Lexicon from Freiburg University Hospital, Germany (41,316 entries) (Weske-Heck et al., 2002)

In addition,

- the English Specialist Lexicon, which is part of the UMLS (96,621 entries, avoiding acronyms and chemical names) (UMLS, 2005),

has also been converted into the Interchange Format. Up until now, 224,351 lexical entries for the biomedical domain, fully encoded with morpho-syntactic features, were collected covering four languages (cf. Table 11.2 for a sample: The first character of the *Mul* field encodes the part-of-speech: N (noun), A (adjective). In case of nouns, c denotes common nouns, m masculine, s singular, n neuter or nominative, depending on the position. For adjectives, f stands for qualitative, p for positive. The character "$-$" indicates that a particular feature does not fit into the language given (e.g. gender in English) or is unspecified for this entry. The number of different lemmas (thus, ignoring ambiguous lexical information for an entry such as, e.g., case) is 105,317 for English, 29,822 for French, 27,480 for German, and 27,093 for

Lng	Typ	Lem	Mul	Frm	Mfr	Prt
FR	B	doigt	Ncms			
EN	T	finger nail	Nc-sn			
SV	B	digital	Afp-sn			
SV	C	Fingeravtryck	Nc-sn			Finger–avtryck
DE	B	Finger	Ncmsn	Fingers	Ncmsg	
DE	C	Fingerfraktur	Ncfsn	Fingerfrakturen	Ncfpn	Finger–frakturen

Table 11.2: Sample of Compiled Lexical Resources (some fields omitted)

Swedish (a total of 189,712, therefore, 1.2 morpho-syntactic variants are given per lexical entry, in average).

11.3 Linking Format Definition

The cross-lingual connection of corresponding entries is the essence of a multilingual dictionary. This operation transforms a set of monolingual lexicons into a multilingual dictionary. Before this operation, the dictionary entries are independent; afterwards, they are organized as clusters of synonyms or translations. Multiple lexical entries, either in the same language or in different languages, are the denotation of the same object in the reality with a common part of speech argument (POS). Typically, *clavicle* in English and *clavicule* in French denote unambiguously the same object (a bone of the pectoral girdle) and they share the same POS: a common noun. The two corresponding entries are candidates to be linked by a translation relation. A similar relation could be defined with the corresponding adjectives, *clavicular* and *claviculaire*. Unfortunately, the process of translating lexical items is not that straightforward, and a couple of cross-lingual phenomena are problematic to capture, especially regarding the different characteristics of case, gender and number in different languages, as well as multiple derivations, e.g. for adjectives, dependent on whether a definite or indefinite object follows or whether their use is attributive or predicative.

Consider the German (Swedish) words *Schere* (*sax*), *Hose* (*bralla*) (both noun,

singular), *Scheren* (*saxar*), *Hosen* (*brallor*) (both noun, plural) and the English equivalents, *scissors* and *trousers* (both noun, plural). Singular forms of the latter examples do not exist,[2] while for other pairs of lexemes, of course, singular forms can be translated to a corresponding singular form in the other language. This information should be kept in a multilingual lexicon, e.g. for the use in machine translation applications.

Different languages also make different use of grammatical gender or noun classes. While in German, Greek or Latin, three grammatical genders are distinguished (masculine, feminine and neuter), French, Portuguese and Spanish only use two (masculine, feminine). Swedish and Danish discriminate the classes *common* and *neuter*. Finally, English does not account for any of these features at all.

In a first version, in order to find an agreement on the question, in which cases two lexical items from different languages, A and B, can be regarded as translations (or, within one language, synonyms) of each other, the following "grades" of bi-directional relationships are defined:

1. **Synonymy/Translation (S/T):** A and B share the same part of speech (POS) and all MULTEXT features, except of gender

2. **Synonymy/Translation, inflected (S/T-i):** A and B share the same POS, but at least one MULTEXT feature differs

3. **Synonymy/Translation, derived (S/T-d):** A and B do not share the same POS

Having these types of relations in mind, a simple Linking Format was created, which is depicted in Table 11.3.

Given this framework, MorphoSaurus is used for the cross-lingual alignment of lexical entities on the semantic level.

[2]except for noun compounds, as evidenced by *"trouser board"* or *"scissor kick"*

Field	Description	Definition
Src	Source Entry ID	ID of the source entry to be linked to a target entry
Tar	Target Entry ID	ID of the target entry linked from the source entry
Typ	Link Type	Type of relation

Table 11.3: Fields of the Linking Format

11.4 Cross-Lingual Alignment

A great deal of work has already been done for the fully automatic cross-lingual alignment of lexical items, most of them using aligned corpora and employing statistical methods, such as context vector comparison (Rapp, 1999; Widdows et al., 2002; Déjean et al., 2002) or mutual information statistics (Fung, 1998). Considering the medical domain, in which multilingual resources are available, e.g. within the UMLS, methods for the automatic search for translation candidates have also already been explored. One promising idea was to use already existing translations at a subword level in order to support the acquisition of translations at a term level (Namer & Baud, 2005; Daumke et al., 2005b). Therefore, the MorphoSaurus system seems particularly well suited for the cross-lingual linkage of available monolingual lexicons.

In a first step, all lexical entries were processed with the morpho-semantic indexing procedure MSI, as described in Section 3.2. After resolving ambiguous MIDs (Chapter 7), a quite simple algorithm was used to perform the mappings between all entries: Every lexeme i and its attributes is compared to any other lexeme j in the list. If their representations in the interlingua format are identical, they are considered as potential translations or synonyms and linked. Then the relation type (S/T, S/T-i, S/T-d, cf. previous section) is determined, by comparing the lexical attributes of the items involved.

11.5 Results

Using the algorithm introduced, 651,542 bi-directional relations between lexemes were obtained, a sample of which is depicted in Table 11.4. For English-German,

Typ	Lng-1	Lem-1	Mul-1	Lng-2	Lem-2	Mul-2
S/T	EN	abdominal hernia	Nc-sn	SV	bukbråck	Nc-sn
S/T-i	EN	abdominal aorta	Nc-sn	DE	Bauchaorten	Ncfpn
S/T-d	EN	alveolar	Afp–n	FR	alvéole	Ncfs

Table 11.4: Sample Links between Lexical Items

126,504 translations were generated (31,544 when only different lemmas are taken into account, thus ignoring ambiguous lexical information), for English-French 70,680 (24,368, respectively) and for English-Swedish 86,655 (34,030). Furthermore, 21,604 (8,312) relations were extracted for French-Swedish, 32,659 (10,458) for French-German and finally, 41,469 (12,105) for German-Swedish. All other relations (271,971) cover intralingual synonymy. The distribution of different types of relations is 66,641 occurrences for S/T (10%), 286,880 for S/T-i (44%) and 298,021 for S/T-d (46%).

11.5.1 Coverage

The UMLS Metathesaurus is the most comprehensive resource for medical terminology. Therefore, it is particularly interesting how many terms of the UMLS are covered by the multilingual lexicon. Table 11.5 (second column) gives the numbers for those items in the UMLS, which are marked as a *preferred entry* and only contain alphabetic characters (thus, multi-word entries and chemical compounds are not considered in the following discussion). Column three gives the number of those UMLS entries, which are covered by the multilingual lexicon. Values range between 13% for German up to 71% for Swedish. The numbers in Column four show how many synonyms and morpho-syntactic variants of UMLS terms are listed in the lexicon which are *not* part of the Metathesaurus, and, therefore, could be added. This consideration only takes those variants into account, which share at least the same part of speech with the corresponding UMLS entry (only S/T and S/T-i).

Finally, the number of additional lexemes in the lexicon that are neither found in the Metathesaurus, nor constitute morpho-syntactic variants of existing UMLS

Language	UMLS	Covered	Synonyms	Additional
English	122,035	32,668	3,807	68,842
German	21,162	2,832	1,269	23,379
French	10,260	3,590	309	25,923
Swedish	12,012	8,520	994	17,579
$\sum$	165,469	189,712		

Table 11.5: Comparison of Lexical Entries: UMLS Metathesaurus and Multilingual Lexicon

entries, are depicted in Column five. In total, the multilingual lexicon contains 189,712 different lemmas, i.e. 24,243 more than the part of the UMLS considered here.

11.5.2 Cross-Lingual Mappings

For the language pairs considered, the UMLS Metathesaurus already contains between 6,700 and 16,000 translations (cf. Table 11.6, Column two). Within a range of 8% (EN-DE and DE-SV) to 36% (EN-SV), these mappings are also included in the multilingual lexicon (Column three). A total of 30,282 synonymous entries (Column four) could be added to 64,837 existing UMLS translations. Finally, those cross-lingual mappings which are captured in the multilingual lexicon but not in the UMLS Metathesaurus, sum up to 81,321 alignments (again, only considering the relations S/T and S/T-i). While there are 64,837 word-to-word translations in the UMLS for the languages considered, the multilingual lexicon contains 120,817 different translations.

11.6 Discussion

In this chapter, a common framework for the integration of heterogeneous lexical resources covering different languages has been introduced. Furthermore, a simple linkage format has been defined, in which lexical relations can be coded. Using such

Language Pair	UMLS	Covered	Synonyms	Additional
English-German	15,979	1,259	8,801	21,484
English-French	12,589	1,783	6,974	15,611
English-Swedish	9,554	3,403	10,124	20,503
German-French	9,859	850	773	8,835
German-Swedish	10,063	810	1,699	9,596
French-Swedish	6,793	1,109	1,911	5,292
$\sum$	64,837	120,817		

Table 11.6: Comparison of Cross-Lingual Mappings

a simple architecture eases the integration of different language pairs.

It has been shown that a substantial amount of subword-based translations can be generated using the MorphoSaurus system. First examinations of the data proved many alignments to be valid (which is also evidenced by those entries and relations that are also part of the UMLS Metathesaurus). Some erroneous translations are due to the coarse-grained semantics underlying the MorphoSaurus model, since it is tailored for text retrieval rather than for exact translations. Many suffixes that encode subtle differences in meaning are ignored in the subword model. This explains, for instance, the automatic alignment of the English word *therapist* to German *Therapie* (*"therapy"*). Obviously, such kind of relation can only be identified by a sophisticated multilingual word model.

The collection of raw lexical data in the medical domain and the identification of translations is an ongoing initiative. An extensive evaluation of the multilingual medical lexicon is still a desideratum.

Chapter 12

Scalability, Generalizability and Limitations of Subword Indexing

A series of proof-of-concept implementations are available in order to show the benefits and scalability of the subword approach with respect to Cross-Language Information Retrieval. It can be shown that the MorphoSaurus approach can be applied for indexing huge document collections in different medical subdisciplines. Furthermore, there is evidence that other technical domains such as law or economics are suited for the adaptation of the subword approach, as well.

12.1 Applications

12.1.1 Searching in Scientific Databases

Institutions such as the U.S. National Library of Medicine manage dozens of databases of medical content, each containing up to 15 million entries (publications, product and pharmaceutical data, etc.). In this context, MorphoSaurus' capability of multilingual document retrieval in no less than six European languages is of special importance with regard to content information in English (e.g. scientific publications).

In collaboration with the German Institute for Medical Documentation and In-

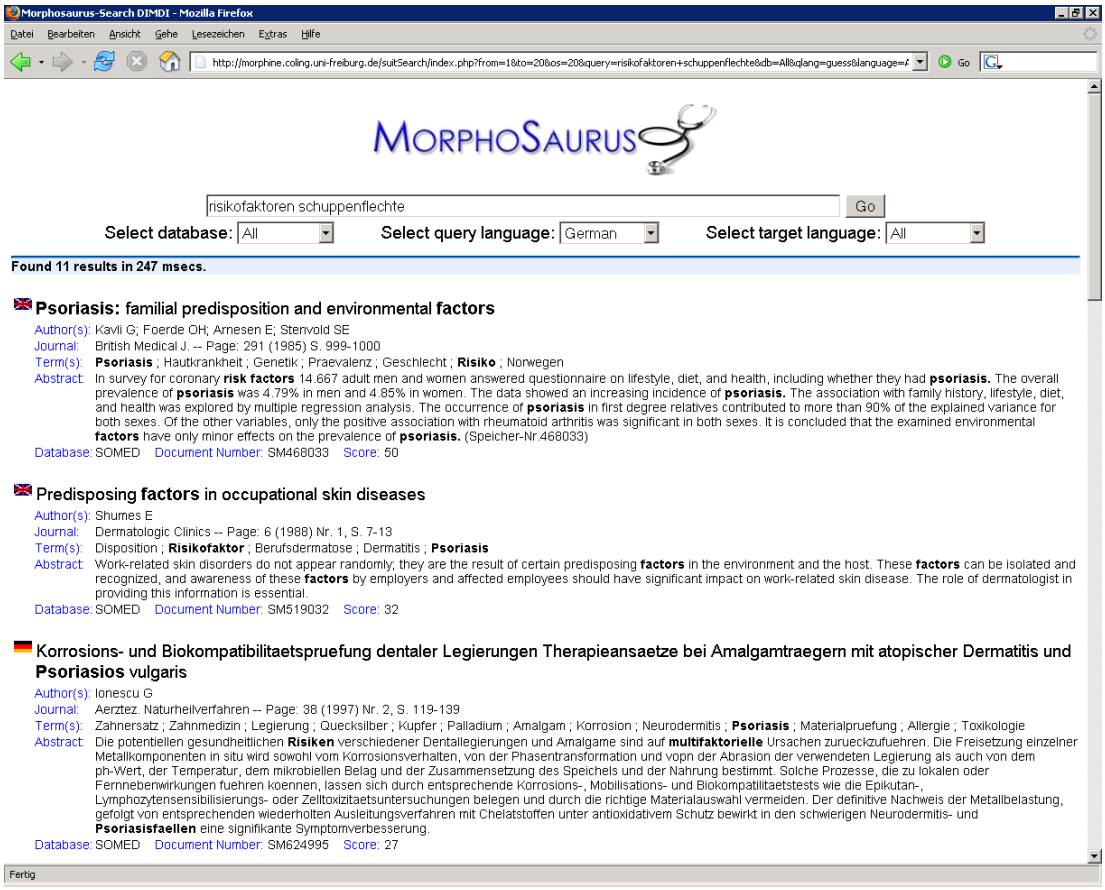

Figure 12.1: Multilingual Bibliographic Information Retrieval

formation Services (DIMDI[1]) MorphoSaurus was implemented for multilingual bibliographic searches. Figure 12.1 depicts the user interface for the search in two heterogeneous databases which are maintained by DIMDI, one covering the fields of social medicine (SOMED), the other focusing on peripheral regions of medicine, such as health policy, health care financing, medical products, etc. (HECLINET: *Health Care Literature Information Network*). A total of more then 650.000 multilingual documents were indexed. The figure illustrates an interface to those databases for which German queries also retrieve documents with synonymous expressions in different languages.

Another showcase application has been implemented for the German Na-

[1] `http://www.dimdi.de/`

tional Library of Medicine (ZBMed[2]). A multilingual search engine based on MorphoSaurus has been made available for searching within the 'Current Contents Medicine' (CCMED) database, a bibliographic repository including more than 1,000 medical journals, which are not accessible via PubMed, the online interface to MEDLINE. It currently covers more then 320,000 references.

Starting in 2007, MorphoSaurus will be installed at ZBMed for providing an intelligent, multilingual search engine for all contents maintained by the institution which sum up to over 240 million articles, including the whole content of MEDLINE. To the best of knowledge, this will give multilingual access to one of the most important (bio-) medical information repositories for the first time.

12.1.2 Searching in Electronic Health Records

In individual healthcare and disease management, the efficient retrieval of documents is a task required on a daily basis. With the introduction of electronic patient files, sophisticated search facilities become increasingly important. According to its simplest definition, the electronic health record (EHR) is a *computer-stored collection of health information about one person linked by a person identifier* (Waegemann, 1996; 2002). On the other hand, the Healthcare Information and Management Systems Society (HIMSS) is claiming more: *The Electronic Health Record (EHR) is a secure, real-time, point-of-care, patient centric information resource for clinicians. [...] The EHR also supports the collection of data for uses other than direct clinical care, such as billing, quality management, outcomes reporting, resource planning, and public health disease surveillance and reporting.*[3]

According to those definitions and to individual patient care, access to the medical information contained in current Hospital Information Systems (HIS) is mostly horizontal, i.e. patient-centered (cf. Figure 12.2). The HIMSS definition suggests more scenarios of use by aggregating information in the vertical view of all electronic patient records. This information relies usually on structured entries like billing information, coded diagnoses and procedures, structured laboratory or microbiology

[2] http://www.zbmed.de/

[3] http://www.himss.org/content/files/ehrattributes070703.pdf

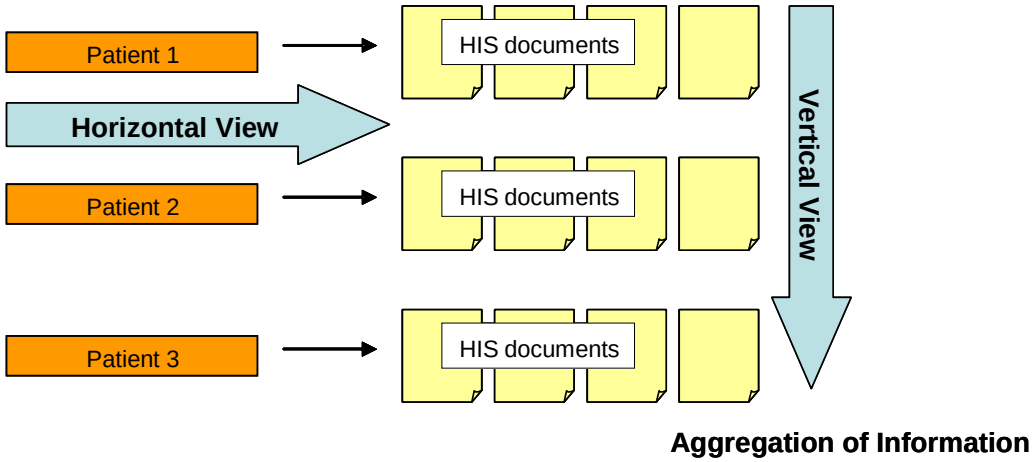

Figure 12.2: Views on the Electronic Health Record

results: it easily can be selected using appropriate and well-known database and data warehouse technologies. On the other hand, for the clinician, non-structured and very heterogeneous information such as admission or discharge summaries and finding reports (pathology, radiology, etc.) and other narrative data are of high relevance for patient care. The more information is stored in the HIS, the more interesting are its vertical, i.e. inter-patient interdependencies.

In conjunction with the Department of Dermatology at the University Hospital in Freiburg (Germany), a search engine for patient reports has been realized employing the MorphoSaurus technology. The user can search the free text portions of the reports for key words in addition to being able to seek out other patient-specific information such as name, patient ID, date of report, authorship etc. Supplementary to any exact matches to a given query, the system also recovers documents containing synonymous information, independent of any linguistic variations that might exist with regard to the query. These new features allow a clinician to pose questions such as:

- *"Which patients did I treat that had the same symptoms?"*
- *"What was the treatment and what was its outcome?"*
- *"Did I treat patients with disease X and symptoms Y?"*

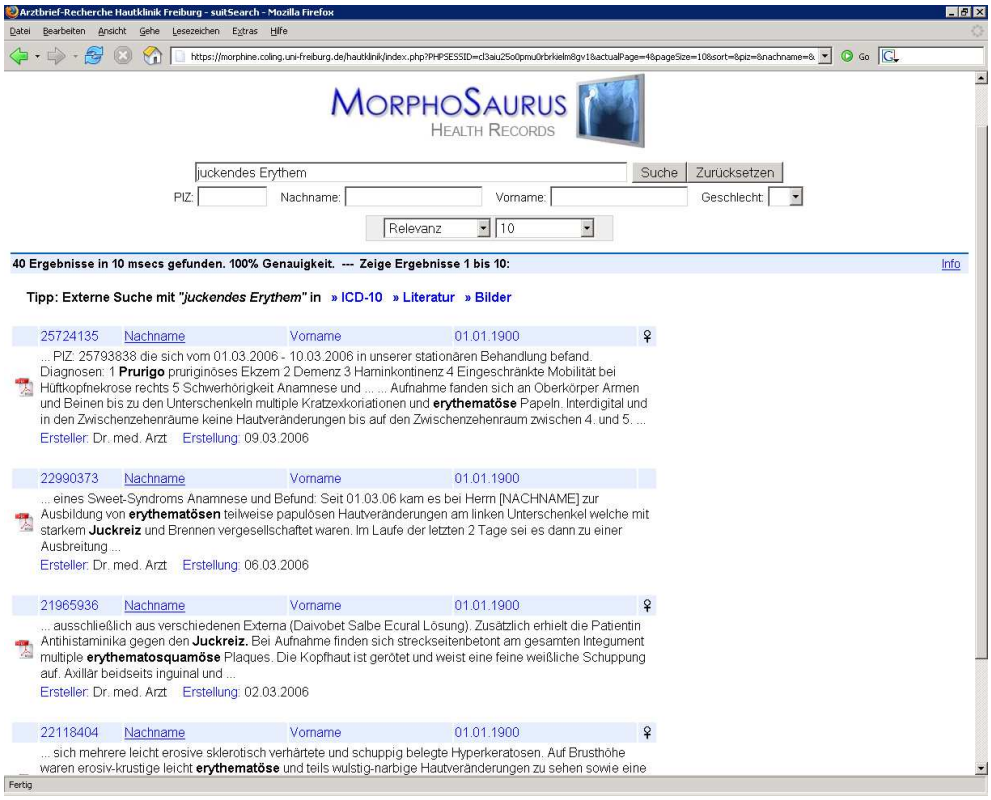

Figure 12.3: MorphoSaurus Search for Electronic Health Records (Anonymized)

- *"Can I have a discharge summary for a patient with disease X that I can use as a template?"*
- *"What was the name of the person with that particular symptom X that I treated three weeks ago?"*

Although several promising technologies like the Clinical Document Architecture (Dolin et al., 2006) and medical terminologies have been developed in order to standardize and structure clinical information, there is still a large gap between this clinical need and today's practice. Here, intelligent search facilities within narrative data, as implemented with the MorphoSaurus system, can augment existing HIS functionality for clinical, scientific, educational and economic reasons. Figure 12.3 shows a screenshot of the application using an anonymized sample of the underlying patient data.

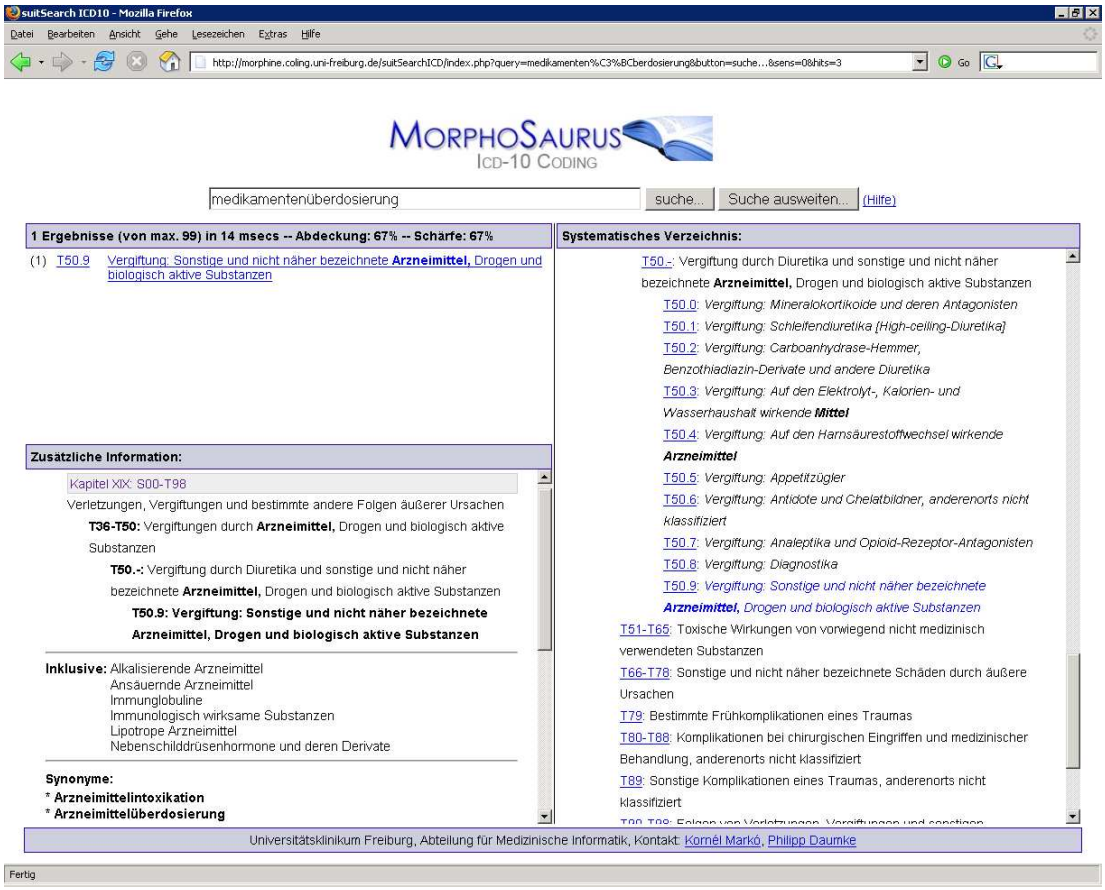

Figure 12.4: ICD Coding based on MorphoSaurus (German)

12.1.3 Searching in Medical Terminology Systems

Another application presented in Figure 12.4 is a coding system for the International Classification of Diseases ICD-10 (2005) based on MorphoSaurus. With the introduction of DRGs (diagnosis related groups) as a performance-oriented and fixed-rate system of financial reimbursements in the health care system, coding of diagnoses and procedures has gained enormous importance in some countries. To this purpose, physicians have to invest significant efforts in the careful assignment of disease and procedure codes. Whereas diseases are globally being encoded by the International Classification of Diseases (ICD), no universal procedure encoding systems exist. In Germany, the classification OPS-301 (OPS, 2006) is used to encode diagnostic and therapeutic procedures, while other countries use different ones,

e.g. CCAM (*Classification Commune des Actes Médicaux*) in France or ICD-9-CM (*clinical modification*) in the United States.

An efficient assignment of medical information to these indexing systems dictates the need for intelligent coding systems. In this, the ability to combine several different ways of accessing the classifications as well as the quality of the test-oriented access (search of key words) decisively influence correctness, quality, and general performance of the coding effort. Here, MorphoSaurus is used to supply the efficient search of key words, mediating between the user's query and the indexing system on the level of subwords. Moreover, foreign physicians and employees, who are not familiar with country-specific classification systems, are given multilingual access to aid in finding the correct codes.

12.1.4 Multimodal Retrieval

As introduced in Section 8.1, the ImageCLEFMed 2006 corpus was used for the evaluation of MorphoSaurus in a cross-lingual environment. At the same time, a search interface to more than 40,000 medical images (mainly covering pathology and radiology for educational purposes) has been implemented. Figure 12.5 shows a screenshot of the application where images with English, French and German captions are retrieved based on a German user query.

Another interface has also been implemented where all search alternatives, i.e. search in health records, bibliographic databases, medical classifications and, finally, pathology and radiology images are accessible within on framework. The user who enters a query can easily switch between the different modalities.

12.2 Generalizability of the Subword Approach

The question that may arise now is whether the subword approach that is proposed in this work can be adopted to other domains, as well. Whenever large, domain-specific (mono- or multilingual) terminologies exist to help people in managing their documents, correspondences or databases by the provision of synonymous terms, this may be the hint for a potential beneficial use of MorphoSaurus. At the

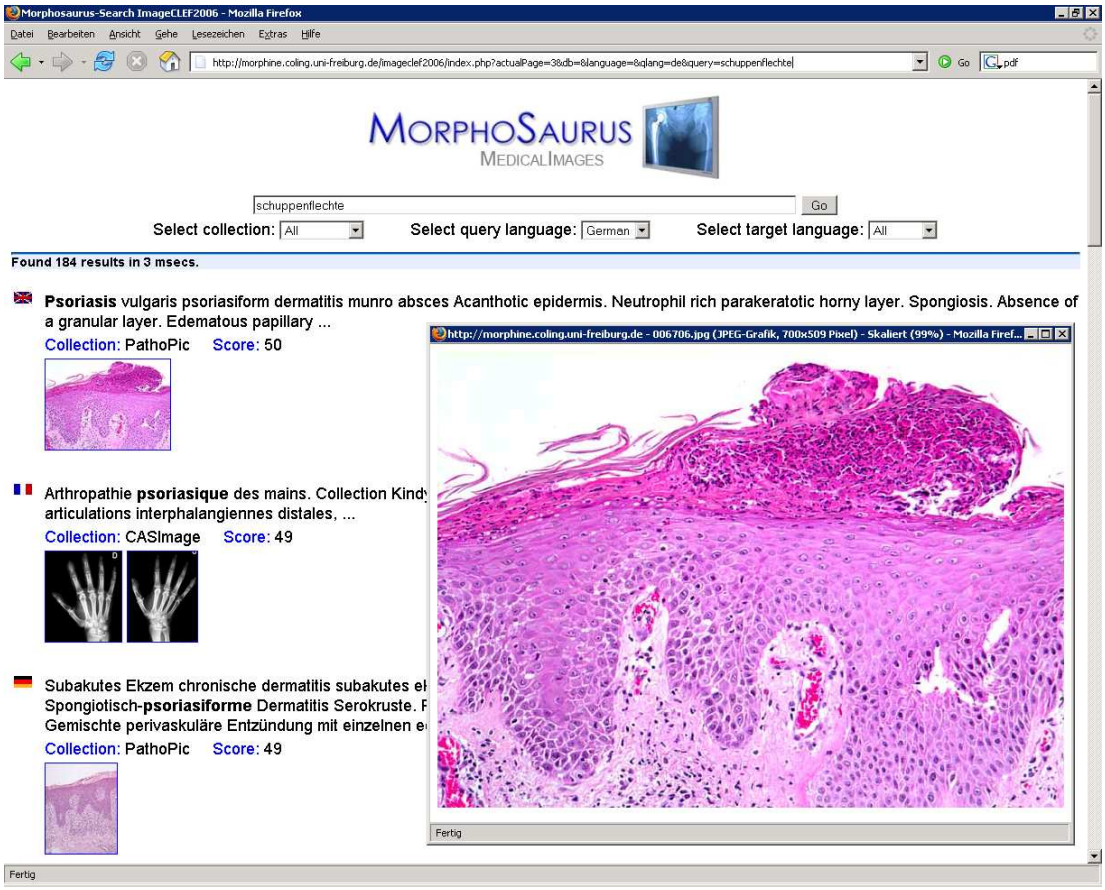

Figure 12.5: Image Retrieval

same time, these reference terminologies can serve as the basis for subword lexicon population for covering these domains.

For the automatic acquisition of lexical entries (cf. Chapter 5) in order to support cross-lingual applications, the availability of large aligned corpora can be regarded as the crucial point. For example, with EUROVOC[4] there exists a thesaurus covering 13 languages in the fields in which the European Community is active, i.e. politics, international relations, law, economics, trade, science, transport, environment, agriculture, education, etc. (cf. Table 12.1). Its entries, as well as those from other terminologies mentioned subsequently, are arranged similarly to those of the UMLS Metathesaurus, including word-to-word translations and complex noun phrases. The

[4] `http://europa.eu.int/celex/eurovoc/`

Thesaurus	Languages	Subject
Eurovoc	13	European Community activities: science, politics, law, culture, economics, etc.
GEMET	19	
UNESCO	3	
OECD	4	
Eurodicautom	12	technical terminology
Europ. Education	18	education, teaching, individual development research, etc.
Europ. Schools Treasury Browser	13	
AGROVOC	6	agriculture
Astronomy Thes.	5	astronomy

Table 12.1: Overview of Selected Multilingual Resources

GEneral Multilingual Environmental Thesaurus (GEMET)[5] covering 6,500 terms in 19 languages, the UNESCO Thesaurus (UNESCO, 1995) covering English, French and Spanish, and the OECD Macrothesaurus[6] (English, German, French, Spanish) all include subject terms for the following areas of knowledge: education, science, culture, social and human sciences, information and communication, politics, law and economics. The Eurodicautom classification[7] includes technical and specialized terminology such as telecommunications, transport and finance in 12 languages. The European Education Thesaurus (EET)[8] as well as the European Treasury Browser Thesaurus[9] focus on terms concerning education, teaching, individual development, etc. in over 11 languages. AGROVOC[10] covers the area of agriculture in English, French, Spanish, Portuguese, Czech, Chinese and Arabic. The Astronomy

[5] http://www.eionet.eu.int/GEMET

[6] http://info.uibk.ac.at/info/oecd-macroth/

[7] http://europa.eu.int/eurodicautom/

[8] http://www.eurydice.org/TeeForm/

[9] http://etb.eun.org/etb/index.html

[10] http://www.fao.org/agrovoc/

Thesaurus[11], to give a last reference, covers English, French, German, Italian and Spanish.

The coverage of these thesauri range from, e.g. 6,500 descriptors translated to 19 languages in GEMET, up to over five million entries (terms and abbreviations) in Eurodicautom.

With the existence of these resources, it has already been shown that there is a need for structuring information in terms of using controlled vocabularies in other domains than medicine. Using subwords as representation units instead of full word forms can substantially reduce the amount of work in organizing those thesauri.

As a conclusion, in what concerns the generalizability of MorphoSaurus and the work presented here, for the proposed lexical acquisition approach on the level of subwords (Chapter 5), one can rely on large-coverage multilingual thesauri available for several relevant domains (cf. Table 12.1), both in terms of the number of languages covered and the number of alignment units available. Acronyms also play a crucial role in other domains than medicine, as evidenced by the high amount of acronym entries in the Eurodicautom thesaurus. The methods for the cross-lingual alignment of acronyms and their expansions (Chapter 6) are useful in understanding how these abbreviations are used in different domains and languages. The methods for cross-lingual disambiguation of subwords (Chapter 7) can be used in a straightforward way. What concerns the evaluation of MorphoSaurus (Chapter 8 and 9) in other domains, one could refer to the GIRT corpus (*German Indexing and Retrieval Testdatabase*, with alignments to English (Kluck, 2004)), which is also used for the CLEF campaign (cf. Section 8.1). It mainly covers the areas of social sciences. The assignment of descriptors from a controlled vocabulary to documents (Chapter 10) is also an important need in, e.g. the industrial domain (for example the *North American Industry Classification System*[12] or the European equivalent *Nomenclature Génerale des Activités Économiques dans les Communautés Européennes*[13]). If products from different manufacturers are assigned one or several key(s) of one of

[11] `http://msowww.anu.edu.au/library/thesaurus/`

[12] `http://www.census.gov/epcd/www/naics.html`

[13] `http://forum.europa.eu.int/irc/dsis/nacecpacon/info/data/en/index.htm`

these classifications, electronic trade across different branches and language barriers is efficiently made possible.

Hence, the MorphoSaurus system and its underlying methodology bears further potential in other domains than Medicine, as well.

12.3 Limitations of the Subword Approach

The subword approach drawn up in this work delineates an efficient way to cover morphological phenomena which notoriously cause so many problems during the processing of natural language expressions, especially with regard to single noun composition. Using high-quality specialized lexicons for the automatic deflection, dederivation and decomposition rarely leads to false segmentations of words within a particular domain, in contradistinction to other systems which are based on heuristic rules or statistical analysis. At the same time, the need of curating such lexical repositories can be seen as the main drawback of the MorphoSaurus system. Lexicographers have to discriminate between subtle shifts of meaning, which immediately have effect on the performance of the system, e.g. for information retrieval or term mapping.

For example, when defining equivalence classes of subwords, fuzzy semantic boundaries may lead to a loss of performance. The term *"somnolent"* can be regarded as a synonym to *"sleepy"* which is derived from *"sleep"*. Grouping together the corresponding subwords into one equivalence class has the effect that querying for *"somnolence"* retrieves any document (or term) in the collection containing *"sleep"*, which in many cases would be undesirable.

Another issue that frequently influence the performance of lexicon-based natural language processing systems is the treatment of so-called out-of-vocabulary words, i.e. terms which cannot be processed due to missing lexicon entries. Within the MorphoSaurus system, words that are not covered by the subword model and the underlying lexicons are restituted and, therefore, available in their original form for further processing. But missing specifications can still lead to severe errors during semantic analysis. As an example, the German word *"Venedig"* (counterpart

for *"Venice"*) gets (formally correct) segmented into *"vene⊕d⊕ig"*, if not specified separately in the lexicon. The German suffixes *"d"* and *"ig"* do not have a particular meaning (and thus, are ignored). The stem *"vene"* (English *"vein"*) is linked to the MID */vein*, together with all other subwords sharing the same meaning.. As a consequence, the German query *"Veneninsuffizienz"* or – in a cross-lingual setting – its translation *"venous insufficiency"* may unintentionally match any document containing the German word *"Venedig"*. A sufficient lexical coverage of a given domain is therefore a prerequisite for applications based upon the subword model.

However, such false segmentations can be avoided by incorporating methods for *Named Entity Recognition* (NER) or even more sophisticated approaches to text understanding, e.g. by incorporating syntactic and ontological knowledge during natural language processing. After syntactic preprocessing (part-of-speech annotation, chunking to phrase groups, parsing) unstructured texts can be semantically enriched by assigning object classes to language expressions. For example, each occurrence of a city or drug name is marked with a special tag which enables a differentiated subsequent processing of those entities. In particular, word and phrase ordering constraints, which are not determined in the bag-of-words approach pursued in this work, can be used to properly interpret, e.g. negated statements or prepositional phrases.

Thus, the conflation of the MorphoSaurus system (that basically operates on the word level) with syntactic analysis (which take effect on the phrase and sentence level), and the integration of additional (ontological) knowledge resources seem to constitute a promising challenge for future work.

Chapter 13

Conclusions

The main goal of this work has been to provide a theory for implementation and evaluation of subword indexing for Cross-Language Information Retrieval and related applications.

Given the productivity of medical terminology it seems almost impossible to create, maintain, and curate high-coverage lexicons, dictionaries and thesauri. The automatic morphological segmentation of words into subwords and their cross-lingual organization in a thesaurus based on these morphological units is one way to face this challenge. A pragmatic approach for defining atomic units (subwords) is used for the automatic deflection, dederivation and decomposition of complex word forms. By grouping subwords into classes of equivalent expressions within (synonymy) and across languages (translation), effective cross-lingual free-text retrieval is made possible, with comparably low manual effort. At the same time, performance increases substantially for mono- as well as multilingual retrieval, as shown in different retrieval settings.

It has also been shown how machine learning algorithms can be used for the acquisition of new subword lexemes for different languages. By using bilingual corpora which are available (not only) for the medical domain, new subwords of a particular language are identified and automatically aligned to already available resources. Similarly, biomedical acronyms and their definitions can be linked across different languages. Furthermore, a new probabilistic methodology for the automatic reso-

lution of multiple word senses has been proposed. It is based upon cross-lingual considerations on the level of subwords. The remarkable impact of subwords disambiguation on retrieval effectiveness is evidenced by large-scale evaluations carried out on standardized test sets. The interlingual representation of textual input is also the basis for the classification according to medical terminologies. In the work at hand, a new approach for the automatic assignment of document descriptors has been elaborated, in which indexing patterns from one language are learned for the benefit of others.

In a series of proof-of-concepts it has been demonstrated that the MorphoSaurus approach can be used in different applications, such as bibliographic search, retrieval within electronic patients records, medical images or medical classifications. The MorphoSaurus technology offers both medical professionals and the general information-seeking public an easy-to-use query interface in order to retrieve health-related content. The importance of this capability is also underlined by market researchers which estimate that about 90% of health care professionals use the Internet for researching clinical matters, reading journal articles (78%) or continuing medical education (45%).[1] Similarly, 80% of all people with Internet access use the Web for searching health information,[2] which is increasingly available in many different languages. In the U.S., each day there are more people seeking medical information on the Web than visiting physicians (Fox & Rainie, 2002). Thus, medical information systems contribute much to the empowerment of health care consumers (Eysenbach, 2000). A partnership on equal terms between health professionals and well informed consumers/patients is becoming more and more accepted.

Considering clinical information systems in intranets, on the other side, the electronic health record is an important challenge in contemporary medicine. It should contain all patients medical information from multiple sources. Since it should be

[1]Taylor, H. & Leitman, R. (2001): The Increasing Impact of eHealth on Physician Behavior (http://www.harrisinteractive.com/news/newsletters/healthnews/HI_HealthCareNews2001Vol1_iss31.pdf)

[2]Taylor, H. (2002): Cyberchondriacs Update.
http://www.harrisinteractive.com/harris_poll/index.asp?PID=299

accessible by any provider caring for the patient, intelligent search facilities have to be provided. Additionally, the medical record should be available from different locations, thus, interoperability has to be guaranteed. For this purpose, the assignment of information in terms of controlled vocabularies (such as MeSH, ICD, etc.) or more sophisticated, domain-specific ontologies is a major desideratum. With MorphoSaurus, there exists a methodology for easing this process, including the possibility to exchange information and knowledge across different languages.

As a conclusion, if the access to health information is one prerequisite for improving the health of society, then the outcome of this work can be regarded as a small, but substantial contribution for reaching this goal.

Chapter 14

Acknowledgments

This work was partly supported by the European Network of Excellence "Semantic Mining" (NoE 507505), by Deutsche Forschungsgemeinschaft (DFG), grant Klar 640/5-1, and by the Brazilian National Council for Scientific Research and Development (CNPq), grants 551277/01-7 and 550240/03-9.

I want to thank Prof. Dr. Udo Hahn (Jena University, Germany) and Prof. Dr. Rüdiger Klar (University Hospital Freiburg, Germany) for accompanying and supporting my work during the last five years. With their different sights on the field of application, computational linguistics on the one hand and medicine on the other, their influence is well mirrored in the interdisciplinary work at hand.

First discussions on the topic of subword indexing were initiated by Stefan Schulz (University Hospital Freiburg, Germany). Together with my colleague Philipp Daumke, who helped with his medical expertise, the theory was well adopted to the medical domain.

The lexicon editing tool was implemented by Edson José Pacheco (Paraná Catholic University Curitiba, Brazil), who was supervised by Percy Nohama, head of Health Technology Master Program at the Brazilian partner university. The lexicographers for the German and English subword lexicons are Claudia Fink, Susanne Hanser, Martin Krüger, Eva Schulte, Martin Schwarz and Oliver Würstlin. The Portuguese lexicon is maintained in Curitiba by Maria Claudia Hahn, Thais Ariela Machado, Josiane Christine Melchioretto and Luciana Bandeira Mendes

Ribeiro. Anders Thurin from Göteborg University Hospital and Mikael Nyström from Linköping University (both Sweden) helped in the development and refinement of the Swedish subword lexicon.

Early experiments with the OHSUMED corpus were assisted by Michael Poprat (Jena University, Germany) and Olena Medelyan (now University of Waikato, New Zealand). Jan Paetzold (Freiburg University Hospital) helped with the preparation and processing of the IMAGECLEFMED data.

The heuristic method for MESH indexing was implemented by Philipp Daumke who also provided the evaluation scripts.

The French resources for the multilingual medical dictionary were supplied by Pierre Zweigenbaum (INSERM, U729, Paris, France) and Robert Baud (University Hospitals of Geneva, Switzerland). Swedish lexicons and additional English resources were made available by Lars Borin (Göteborg University) and Magnus Merkel, Linköping University.

For proof-reading this manuscript I want to thank Stefan Schulz and Martin Romacker (Novartis Pharma AG Basel, Switzerland).

And last, but certainly not least, I thank my parents and family for always supporting me in everything I do.

Bibliography

Adar, Eytan (2004). SaRAD: A simple and robust abbreviation dictionary. *Bioinformatics*, 20(4):527–533.

Allen, James (1995). *Natural Language Understanding* (2nd ed.). Redwood City, CA: Benjamin/Cummings.

Aronson, A. R., O. Bodenreider, H. F. Chang, S. M. Humphrey, J. G. Mork, S. J. Nelson, T. C. Rindflesch & W. John Wilbur (1999). The indexing initiative. In *A Report to the Board of Scientific Counselors of the Lister Hill National Center for Biomedical Communications*. Bethesda, MD: National Library of Medicine.

Aronson, Alan R. (2001). Effective mapping of biomedical text to the Umls Metathesaurus: The MetaMap program. In Suzanne Bakken (Ed.), *AMIA 2001 – Proceedings of the Annual Symposium of the American Medical Informatics Association. A Medical Informatics Odyssey: Visions of the Future and Lessons from the Past*, pp. 17–21. Washington, D.C., November 3-7, 2001. Philadelphia, PA: Hanley & Belfus.

Aronson, Alan R., O. Bodenreider, H. F. Chang, S. M. Humphrey, J. G. Mork, S. J. Nelson, T. C. Rindflesch & W. J. Wilbur (2000). The NLM indexing initiative. In J. Marc Overhage (Ed.), *AMIA 2000 – Proceedings of the Annual Symposium of the American Medical Informatics Association. Converging Information, Technology, and Health Care*, pp. 17–21. Los Angeles, CA, November 4-8, 2000. Philadelphia, PA: Hanley & Belfus.

Aronson, Alan R., James G. Mork, Clifford W. Gay, Susanne M. Humphrey & Willie J. Rogers (2004). The NLM indexing initiative's medical text indexer. In Marius Fieschi, Enrico Coiera & Yu-Chan Jack Li (Eds.), *MEDINFO 2004 – Proceedings of the 11th World Congress on Medical Informatics. Vol. 1*, Studies in Health Technology and Informatics 107, pp. 268–272. San Francisco, CA, USA, September 7-11, 2004. Amsterdam: IOS Press.

Baeza-Yates, Ricardo & Berthier Ribeiro-Neto (Eds.) (1999). *Modern Information Retrieval*. Reading, MA: Addison-Wesley & Longman.

Barker, Gosia & Richard F. E. Sutcliffe (2000). An experiment in the semi-automatic identification of false-cognates between English and Polish. In *AICS 2000 – Irish Conference on Artificial Intelligence and Cognitive Science*. National University of Ireland Galway, 24-25 August, 2000.

Baud, Robert, Mikaël Nyström, Lars Borin, Robert Evans, Stefan Schulz & Pierre Zweigenbaum (2005). Interchanging lexical information for a multilingual dictionary. In *AMIA 2005 – Proceedings of the Annual Symposium of the American Medical Informatics Association*, pp. 31–35.

Baud, Robert H., Christian Lovis, Anne-Marie Rassinoux & Jean-Raoul Scherrer (1998). Morpho-semantic parsing of medical expressions. In C. G. Chute (Ed.), *AMIA'98 – Proceedings of the 1998 AMIA Annual Fall Symposium. A Paradigm Shift in Health Care Information Systems: Clinical Infrastructures for the 21st Century*, pp. 760–764. Orlando, FL, November 7-11, 1998. Philadelphia, PA: Hanley & Belfus.

Baud, Robert H., Anne-Marie Rassinoux, Patrick Ruch, Christian Lovis & Jean-Raoul Scherrer (1999). The power and limits of a rule-based morpho-syntactic parser. In N. M. Lorenzi (Ed.), *AMIA'99 – Proceedings of the 1999 Annual Symposium of the American Medical Informatics Association. Transforming Health Care through Informatics: Cornerstones for a New Information Management Paradigm*, pp. 22–26. Washington, D.C., November 6-10, 1999. Philadelphia, PA: Hanley & Belfus.

Black, Alan W., Joke van de Plassche & Briony William (1991). Analysis of unknown words through morphological decomposition. In *EACL'91 – Proceedings of the 5th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 101–106. Berlin, Germany, 9-11 April 1991. Association for Computational Linguistics.

Blaschke, C., M. A. Andrade, C. Ouzounis & A. Valencia (1999). Automatic extraction of biological information from scientific text: Protein-protein interactions. In *ISMB'99 – Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology*, pp. 60–67. Heidelberg, Germany, August 6-10, 1999. Menlo Park, CA: AAAI Press.

Bodenreider, Olivier, Anita Burgun & Joyce A. Mitchell (2003). Evaluation of WordNet as a source of lay knowledge for molecular biology and genetic diseases: A feasibility study. In Robert Baud, Marius Fieschi, Pierre Le Beux & Patrick Ruch (Eds.), *Medical Informatics Europe 2003 – Proceedings of the 18th International Congress of the European Federation for Medical Informatics. The New Navigators: From Professionals to Patients.*, Studies in Health Technolgy and Informatics 95, pp. 379–384. St. Malo, France, May 4-7, 2003. Amsterdam: IOS Press.

Braschler, Martin, Anne Göhring & Peter Schäuble (2003). Eurospider at CLEF 2002. In *CLEF '02: Revised Papers from the Third Workshop of the Cross-Language Evaluation Forum*, pp. 164–174. Rome, Italy, September 2002. Springer-Verlag.

Braschler, Martin & Bärbel Ripplinger (2004). How effective is stemming and decompounding for German text retrieval? *Information Retrieval*, 7(3-4):291–316.

Braschler, Martin & Peter Schäuble (2000). Experiments with the Eurospider retrieval system for CLEF 2000. In *CLEF '00: Revised Papers from the Workshop of Cross-Language Evaluation Forum on Cross-Language Information Retrieval and Evaluation*, pp. 140–148. Springer-Verlag.

Brigl, Birgit, Markus Mieth, Reinhold Haux & Ewald Glück (1994). The LBI-method for automated indexing of diagnoses by using SNOMED. Part 1: Design and realization. *International Journal of Bio-Medical Computing*, 37(6):237–247.

Brown, Peter F., John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer & Paul S. Roossin (1990). A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

Brown, Peter F., Stephen A. Della Pietra, Vincent J. Della Pietra & Robert L. Mercer (1991). Word-sense disambiguation using statistical methods. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pp. 264–270. Berkeley, CA, USA, 18-21 June 1991. Association for Computational Linguistics.

Bryden, John (2003). Have over 35 years of health informatics made Europe healthier? *British Journal of Healthcare Computing & Information Management*, 20(7):15–17.

Burgun, Anita & Olivier Bodenreider (2001). Comparing terms, concepts and semantic classes in WORDNET and the *Unified Medical Language System*. In *Proceedings of the NAACL 2001 Workshop 'WORDNET and Other Lexical Resources: Applications, Extensions and Customizations'*, pp. 77–82. Pittsburgh, PA, June 3-4, 2001. New Brunswick, NJ: Association for Computational Linguistics.

Campbell, David A. & Stephen B. Johnson (2001). Comparing syntactic complexity in medical and non-medical corpora. In Suzanne Bakken (Ed.), *AMIA 2001 – Proceedings of the Annual Symposium of the American Medical Informatics Association. A Medical Informatics Odyssey: Visions of the Future and Lessons from the Past*, pp. 90–94. Washington, D.C., November 3-7, 2001. Philadelphia, PA: Hanley & Belfus.

Chalmers, I. (1993). The Cochrane collaboration: Preparing, maintaining, and disseminating systematic reviews of the effects of health care. *Annals of the New York Academy of Sciences*, 703:156–163.

Chang, Jeffrey T., Hinrich Schütze & Russ B. Altman (2002). Creating an online dictionary of abbreviations from MEDLINE. *Journal of the American Medical Informatics Association*, 9(6):612–620.

Chen, Aitao (2002). Cross-language retrieval experiments at CLEF 2002. In *Advances in Cross-Language Information Retrieval*, Vol. 2785 / 2003, Lecture Notes in Computer Science, pp. 28–48. Heidelberg, Germany: Springer Verlag.

Cheng, Pu-Jen, Jei-Wen Teng, Ruei-Cheng Chen, Jenq-Haur Wang, Wen-Hsiang Lu & Lee-Feng Chien (2004). Translating unknown queries with web corpora for cross-language information retrieval. In *SIGIR '04: Proceedings of the 27th annual international conference on Research and development in information retrieval*, pp. 146–153. New York, NY, USA: ACM Press.

Chiao, Y.C. & P. Zweigenbaum (2002). Looking for French-English translations in comparable medical corpora. In Isaac S. Kohane (Ed.), *AMIA 2002 – Proceedings of the Annual Symposium of the American Medical Informatics Association. Biomedical Informatics: One Discipline*, pp. 150–154. San Antonio, TX, November 9-13, 2002. Philadelphia, PA: Hanley & Belfus.

Chodorow, Martin, Claudia Leacock & George A. Miller (2000). A topical/local classifier for word sense identification. *Computers and the Humanities*, 34(1/2):115–120.

Choueka, Yaacov (1990). RESPONSA: An operational full-text retrieval system with linguistic components for large corpora. In A. Zampolli, L. Cignoni & E. C. Peters (Eds.), *Computational Lexicology and Lexicography. Special Issue Dedicated to Bernard Quemada. Vol. 1*, Vol. 6, Linguistica Computazionale, pp. 181–217. Pisa: Giardini Editori E. Stampatori.

Ciaramita, Massimiliano, Thomas Hofmann & Mark Johnson (2003). Hierarchical semantic classification: Word sense disambiguation with world knowledge. In *IJCAI'03 – Proceedings of the 18th International Joint Conference on Artificial Intelligence*, pp. 817–822. Acapulco, Mexico, August 9-15, 2003. San Francisco, CA: Morgan Kaufmann.

Clough, P., H. Müller, T. Deselaers, M. Grubinger, Jensen J. Lehmann & W. Hersh (2005). The CLEF 2005 cross-language image retrieval track. In C. Peters, F.C. Gey, J. Gonzalo, H. Müller, G.J.F. Jones, M. Kluck, B. Magnini & M. de Rijke (Eds.), *6th Workshop of the Cross-Language Evaluation Forum, CLEF 2005*. Springer Lecture Notes in Computer Science.

Côté, Roger, David J. Rothwell, Ronald S. Beckett, James L. Palotay & Louise Brochu (1993). *The Systemised Nomenclature of Medicine:* SNOMED *International*. Northfield, IL: College of American Pathologists.

CT, SNOMED (2004). SNOMED *Clinical Terms*. Northfield, IL: College of American Pathologists.

Dagan, Ido & Alon Itai (1994). Word sense disambiguation using a second language monolingual corpus. *Computational Linguistics*, 20(4):563–596.

Dagan, Ido, Alon Itai & Ulrike Schwall (1991). Two languages are more informative than one. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pp. 130–137. Berkeley, CA, USA, 18-21 June 1991. Association for Computational Linguistics.

Daraselia, Nikolai, Sergei Egorov, Andrey Yazhuk, Svetlana Novichkova, Anton Yuryev & Ilya Mazo (2004). Extracting protein function information from MEDLINE using a full-sentence parser. In *Proceedings of the 2nd European ECML/PKDD 2004 Workshop on Data Mining and Text Mining in Bioinformatics*, pp. 11–18. Pisa, Italy, 24 September 2004.

Daumke, Philipp, Stefan Schulz & Kornél Markó (2005a). A CLIR interface to a Web search engine. In *SIGIR 2005 – Proceedings of the 28th Annual Interna-*

tional ACM SIGIR Conference on Research and Development in Information Retrieval. Salvador, Brasil, August 15-19, 2005.

Daumke, Philipp, Stefan Schulz & Kornél Markó (2005b). Searching multilingual medical content in the Web. *Technology and Health Care*, 13(5).

Deerwester, Scott, Susan Dumais, George Furnas, Thomas Landauer & Richard Harshman (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Déjean, Hervé, Éric Gaussier & Fatiha Sadat (2002). An approach based on multilingual thesauri and model combination for bilingual lexicon extraction. In *COLING 2002 – Proceedings of the 19th International Conference on Computational Linguistics*, pp. 218–224. Taipei, Taiwan, August 24 - September 1, 2002. Association for Computational Linguistics.

Dolin, R.H., L. Alschuler, S. Boyer, C. Beebe, F.M. Behlen, P.V. Biron & A. Shabo Shvo (2006). HL7 clinical document architecture, release 2. *Journal of the American Medical Informatics Association*, 13(1):30–39.

Dujols, P., P. Aubas, C. Baylon & F. Grémy (1991). Morphosemantic analysis and translation of medical compound terms. *Methods of Information in Medicine*, 30(1):30–35.

Eco, Umberto, Klarus Robering, Adelhard Scheffczyk & Rainer Habermeier (1988). Metamorphoses of the semiotic triangle. *Zeitschrift für Semiotik*, 10(3).

Eichmann, David, Miguel E. Ruiz & Padmini Srinivasan (1998). Cross-language information retrieval with the UMLS Metathesaurus. In W. B. Croft, A. Moffat, C. J. van Rijsbergen, R. Wilkinson & J. Zobel (Eds.), *SIGIR'98 – Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 72–80. Melbourne, Australia, August 24-28, 1998. New York, NY: ACM.

Ekmekçioglu, F. Cuna, Michael F. Lynch & Peter Willett (1995). Development and evaluation of conflation techniques for the implementation of a document

retrieval system for Turkish text databases. *New Review of Document and Text Management*, 1(1):131–146.

Eysenbach, G. (2000). Consumer health informatics. *British Medical Journal*, 320:1713–1716.

Feldman, Ronen, Yonatan Aumann, Moshe Fresko, Orly Lipshtat, Binyamin Rosenfeld & Yonatan Schler (1999). Text mining via information extraction. In J. M. Zytkow & J. Rauch (Eds.), *Principles of Data Mining and Knowledge Discovery. Proceedings of the 3rd European Conference – PKDD'99*, Vol. 1704, Lecture Notes in Artificial Intelligence, pp. 165–173. Prague, Czech Republic, September 15-18, 1999. Berlin: Springer.

Fellbaum, Christiane (Ed.) (1998). WordNet*: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Ferber, Reginald (1997). Automated indexing with thesaurus descriptors: A co-occurrence based approach to multilingual retrieval. In Carol Peters & Costantino Thanos (Eds.), *Research and Advanced Technology for Digital Libraries. Proceedings of the 1st European Conference – ECDL'97*, Lecture Notes in Computer Science 1324, pp. 232–255. Pisa, Italy, 1-3 September, 1997. Berlin, Heidelberg, New York: Springer.

Fox, S. & L. Rainie (2002). E-patients and the online health care revolution. *Physician Executive*, 28:14–17.

Friedman, Carol, Philip O. Alderson, John H. M. Austin, James J. Cimino & Stephen B. Johnson (1994). A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association*, 1(2):161–174.

Friedman, Carol & George Hripcsak (1998). Evaluating natural language processors in the clinical domain. *Methods of Information in Medicine*, 37(4-5):334–344.

Friedman, Carol & George Hripcsak (1999). Natural language processing and its future in medicine. *Academic Medicine*, 74(8):890–895.

Fuhr, Norbert (1992). Probabilistic models in information retrieval. *Computer Journal*, 35(3):243–255.

Fukuda, F., T. Tsunoda, A. Tamura & T. Takagi (1998). Toward information extraction: Identifying protein names from biological papers. In Russ B. Altman, A. Keith Dunker, Lawrence Hunter & Teri E. Klein (Eds.), *PSB 98 – Proceedings of the 3rd Pacific Symposium on Biocomputing*, pp. 705–716. Maui, Hawaii, USA, 4-9 January, 1998. Singapore: World Scientific Publishing.

Fung, Pascale (1998). A statistical view on bilingual lexicon extraction: From parallel corpora to non-parallel corpora. In David Farwell, Laurie Gerber & Eduard H. Hovy (Eds.), *Machine Translation and the Information Soup. Proceedings of the 3rd Conference of the Association for Machine Translation in the Americas – AMTA 98*, Vol. 1529, Lecture Notes in Computer Science, pp. 1–17. Langhorne, PA, USA, October 28-31, 1998. Berlin: Springer.

Funk, Mark E. & Carolyn Anne Reid (1983). Indexing consistency in MEDLINE. *Bulletin of the Medical Library Association*, 71(2):176–183.

Gale, William A., Kenneth W. Church & David Yarowsky (1993). A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26(5):415–439.

Gay, Clifford W., Mehmet Kayaalp & Alan R. Aronson (2005). Semi-automatic indexing of full text biomedical articles. In *AMIA'05 – Proceedings of the 2005 Annual Symposium of the American Medical Informatics Association*, pp. 271–275. Washington, D.C., November 22-26, 2003. Philadelphia, PA: Hanley & Belfus.

Gey, F. & A. Chen (2000). Trec-9 cross-language information retrieval (English-Chinese) overview. In *Proceedings of the Ninth Text REtrieval Conference (TREC 9)*. NIST Special Publication, No. 500-249.

Gey, Frederic C., Noriko Kando & Carol Peters (2002). Cross-language information retrieval: A research roadmap. *SIGIR Forum*, 36(1):72–80.

Gey, Fredric C.and Kando, Noriko & Carol Peters (2005). Cross-language information retrieval: the way ahead. *Information Processing and Management: an International Journal*, 41(3):415–431.

Goldsmith, John (2001). Unsupervised learning of the morphology of a natural language. *Computational Linguistics*, 27(2):153–193.

Gonzalo, Julio, Felisa Verdejo & Irina Chugur (1999). Using EuroWordNet in a concept-based approach to cross-language text retrieval. *Applied Artificial Intelligence*, 13(7):647–678.

Gospodnetic, Otis & Erik Hatcher (2004). *Lucene in Action*. Manning Publications.

Grefenstette, Gregory (Ed.) (1998). *Cross-Language Information Retrieval*, Vol. 2. Kluwer International Series on Information Retrieval. Boston: Kluwer.

Grefenstette, Gregory & Pasi Tapanainen (1994). What is a word, what is a sentence? problems of tokenization. In *Proceedings of the 3rd International Conference on Computational Lexicography*, pp. 79–87.

Hahn, Udo, Philipp Daumke, Stefan Schulz & Kornél Markó (2005a). Cross-language mining for acronyms and their completions from the Web. In *DS 2005 – Proceedings of the 8th International Conference on Discovery Science*. Singapore, October 8-11, 2005.

Hahn, Udo, Martin Honeck, Michael Piotrowski & Stefan Schulz (2001). Subword segmentation: Leveling out morphological variations for medical document retrieval. In Suzanne Bakken (Ed.), *AMIA 2001 – Proceedings of the Annual Symposium of the American Medical Informatics Association. A Medical Informatics Odyssey: Visions of the Future and Lessons from the Past*, pp. 229–233. Washington, D.C., November 3-7, 2001. Philadelphia, PA: Hanley & Belfus.

Hahn, Udo, Kornél Markó, Michael Poprat, Stefan Schulz, Joachim Wermter & Percy Nohama (2004a). Crossing languages in text retrieval via an interlingua. In *RIAO 2004 – Conference Proceedings: Coupling Approaches, Coupling*

Media and Coupling Languages for Information Retrieval, pp. 100–115. Avignon, France, 26-28 April 2004. Paris: Centre de Hautes Etudes Internationales d'Informatique Documentaire (CID).

Hahn, Udo, Kornél Markó & Stefan Schulz (2004b). Learning indexing patterns from one language for the benefit of others. In *AAAI'04 – Proceedings of the 19th National Conference on Artificial Intelligence & IAAI'04 – Proceedings of the 16th Innovative Applications of Artificial Intelligence Conference*, pp. 406–411. San José, CA, USA, July 25-29, 2004. Menlo Park, CA; Cambridge, MA: AAAI Press & MIT Press.

Hahn, Udo, Kornél Markó & Stefan Schulz (2005b). Subword clusters as light-weight interlingua for multilingual document retrieval. In *MT Summit X – Proceedings of the 10th Machine Translation Summit of the International Association for Machine Translation*. Phuket, Thailand, September 12-16, 2005.

Hahn, Udo, Martin Romacker & Stefan Schulz (2000). MEDSYNDIKATE: Design considerations for an ontology-based medical text understanding system. In J. Marc Overhage (Ed.), *AMIA 2000 – Proceedings of the Annual Symposium of the American Medical Informatics Association. Converging Information, Technology, and Health Care*, pp. 330–334. Los Angeles, CA, November 4-8, 2000. Philadelphia, PA: Hanley & Belfus.

Hahn, Udo, Martin Romacker & Stefan Schulz (2002a). Creating knowledge repositories from biomedical reports: The MEDSYNDIKATE text mining system. In Russ B. Altman, A. Keith Dunker, Lawrence Hunter, Kevin Lauderdale & Teri E. Klein (Eds.), *PSB 2002 – Proceedings of the Pacific Symposium on Biocomputing 2002*, pp. 338–349. Kauai, Hawaii, USA, January 3-7, 2002. Singapore: World Scientific Publishing.

Hahn, Udo, Martin Romacker & Stefan Schulz (2002b). MEDSYNDIKATE: A natural language system for the extraction of medical information from finding reports. *International Journal of Medical Informatics*, 67(1/3):63–74.

Hahn, Udo & Joachim Wermter (2004). High-performance tagging on medical texts. In *COLING Geneva 2004 – Proceedings of the 20th International Conference on Computational Linguistics*, Vol. 2, pp. 973–979. Geneva, Switzerland, August 23-27, 2004. Association for Computational Linguistics.

Harman, Donna (1991). How effective is suffixing? *Journal of the American Society for Information Science*, 42(1):7–15.

Hedlund, Turid, Ari Pirkola & Kalervo Järvelin (2001). Aspects of Swedish morphology and semantics from the perspective of mono- and cross-language retrieval. *Information Processing & Management*, 37(1):147–161.

Hersh, William R. (2002). *Information Retrieval. A Health and Biomedical Perspective* (2nd ed.). New York: Springer.

Hersh, William R., Chris Buckley, T. J. Leone & David H. Hickam (1994a). OHSUMED: An interactive retrieval evaluation and new large test collection for research. In Bruce Croft & C. J. van Rijsbergen (Eds.), *SIGIR'94 – Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 192–201. Dublin, Ireland, 3-6 July 1994. London: Springer.

Hersh, William R. & Larry C. Donohoe (1998). SAPHIRE International: A tool for cross-language information retrieval. In C. G. Chute (Ed.), *AMIA'98 – Proceedings of the 1998 AMIA Annual Fall Symposium. A Paradigm Shift in Health Care Information Systems: Clinical Infrastructures for the 21st Century*, pp. 673–677. Orlando, FL, November 7-11, 1998. Philadelphia, PA: Hanley & Belfus.

Hersh, William R., David H. Hickman, Brian Haynes & K. Ann McKibbon (1994b). A performance and failure analysis of SAPHIRE with a MEDLINE test collection. *Journal of the American Medical Informatics Association*, 1(1):51–60.

Hirst, Graeme (2004). Ontologies and the lexicon. In Steffen Staab & Rudi Studer

(Eds.), *Handbook on Ontologies*, International Handbooks on Information Systems, pp. 209–229. Berlin: Springer.

Hooper, R.S. (1965). *Indexer Consistency Tests: Origin, Measurement, Results, and Utilization*. Bethesda, MD: IBM Corporation.

Hull, David A. (1996). Stemming algorithms: A case study for detailed evaluation. *Journal of the American Society for Information Science*, 47(1):70–84.

Hunter, Lawrence & K. Bretonnel Cohen (2006). Biomedical language processing: Perspective what's beyond PubMed? *Molecular Cell*, 21(5):589–594.

ICD-10 (2005). *International Statistical Classification of Diseases and Health Related Problems. 10th Revision*. Geneva: World Health Organization.

ICPC (1990). *International Classification of Primary Care*. Oxford University Press.

Ide, Nancy (2000). Cross-lingual sense determination: Can it work? *Computers and the Humanities*, 34(1/2):223–234.

Ide, Nancy & Jean Véronis (1998). Introduction to the Special Issue on Word Sense Disambiguation: The state of the art. *Computational Linguistics*, 24(1):1–40.

Ingenerf, Joseph (1997). *Medizinische Linguistik*. Seelos.

Jain, Nilesh L. & Carol Friedman (1997). Identification of findings suspicious for breast cancer based on natural language processing of mammogram reports. In Daniel R. Masys (Ed.), *AMIA'97 – Proceedings of the 1997 AMIA Annual Fall Symposium. The Emergence of 'Internetable' Health Care: Systems that Really Work*, pp. 829–833. Nashville, TN, October 25-29, 1997. Philadelphia, PA: Hanley & Belfus.

Jain, Nilesh L., Charles A. Knirsch, Carol Friedman & George Hripcsak (1996). Identification of suspected tuberculosis patients based on natural language processing of chest radiograph reports. In James J. Cimino (Ed.), *AMIA'96 - Proc. 20th AMIA Annual Fall Symposium (formerly SCAMC); Washington, D.C., 26–30 Oct, 1996*, pp. 542–546. Philadelphia, PA: Hanley & Belfus.

Jäppinen, Harri & Juha Niemistö (1988). Inflections and compounds: Some linguistic problems for automatic indexing. In *RIAO 88 – Proceedings of the RIAO 88 Conference: User-Oriented Content-Based Text and Image Handling*, Vol. 1, pp. 333–342. Cambridge, MA, March 21-24, 1988. Paris: Centre de Hautes Etudes Internationales d'Informatique Documentaire (CID).

Kamps, Jaap, Christof Monz, Maarten de Rijke & Börkur Sigurbjörnsson (2003). Language-dependent and language-independent approaches to cross-lingual text retrieval. In Carol Peters, Julio Gonzalo, Martin Braschler & Michael Kluck (Eds.), *Comparative Evaluation of Multilingual Information Access Systems: 4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003, Trondheim, Norway, August 21-22, 2003*, Vol. 3237, pp. 152–165.

Kantrowitz, Mark, Behrang Mohit & Vibhu Mittal (2000). Stemming and its effects on *tfidf* ranking. In N. J. Belkin, P. Ingwersen & M.-K. Leong (Eds.), *SIGIR 2000 – Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 357–359. Athens, Greece, July 24-28, 2000. New York, NY: ACM.

Kaplan, Abraham (1955). An experimental study of ambiguity and context. *Mechanical Translation*, 2(2):39–46.

Karttunen, Lauri, Ronald M. Kaplan & Annie Zaenen (1992). Two-level morphology with composition. In *COLING'92 – Proceedings of the 14th International Conference on Computational Linguistics*, Vol. 1: Topical Papers, pp. 141–148. Nantes, France, 23-28 August 1992. ICCL.

Kay, Martin (1980). Morphological analysis. In A. Zampolli & N. Calzolari (Eds.), *Computational and Mathematical Linguistics. Proceedings of the International Conference on Computational Linguistics*, Vol. 1, pp. 205–223. Pisa, 27 August - 1 September 1973. Firenze: L. S. Olschki.

Kilgarriff, Adam & Martha Palmer (2000). Introduction to the special issue on SensEval. *Computers and the Humanities*, 34(1/2):1–13.

Kluck, Michael (2004). The GIRT data in the evaluation of CLIR systems from 1997 until 2003. In C. Peters, J. Gonzalo, M. Braschler & M. Kluck (Eds.), *4th Workshop of the Cross-Language Evaluation Forum, CLEF 2003: Comparative Evaluation of Multilingual Information Access Systems*. Springer Lecture Notes in Computer Science.

Koehn, Philipp & Kevin Knight (2002). Learning a translation lexicon from monolingual corpora. In *Unsupervised Lexical Acquisition. Proceedings of the Workshop of the ACL Special Interest Group on the Lexicon (SIGLEX)*, pp. 9–16. Philadelphia, PA, USA, July 12, 2002. Association for Computational Linguistics.

Kokkinakis, Dimitrios & Dana Dannélls (2006). Recognizing acronyms and their definitions in Swedish medical texts. In *LREC 2006 – Proceedings of the 5th International Conference on Language Resources and Evaluation*, pp. 1971–1974. Genua, Italy, May 24-26, 2006.

Kornai, A.and Stone, L. (2004). Automatic translation to controlled medical vocabularies. *Innovations in Intelligent Systems and Applications*, pp. 413–434.

Koskenniemi, Kimmo (1984). A general computational model for word formation recognition and production. In *COLING'84 – Proceedings of the 10th International Conference on Computational Linguistics & 22nd Annual Meeting of the Association for Computational Linguistics*, pp. 178–181. Stanford, California, U.S.A., 2-6 July 1984.

Krovetz, Robert (1993). Viewing morphology as an inference process. In R. Korfhage, E. Rasmussen & P. Willett (Eds.), *SIGIR'93 – Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 191–203. Pittsburgh, PA, USA, June 27 - July 1, 1993. New York, NY: ACM.

Leacock, Claudia, Geoffrey Towell & Ellen M. Voorhees (1996). Towards building contextual representations of word senses using statistical models. In Branimir

Boguraev & James Pustejovsky (Eds.), *Corpus Processing for Lexical Acquisition*, pp. 97–113. Cambridge, MA: MIT Press.

Lee, Yoong Keok & Hwee Tou Ng (2002). An empirical evaluation of knowledge sources and learning algorithms for word sense disambiguation. In *EMNLP'02 – Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pp. 41–48. University of Pennsylvania, Philadelphia, PA, USA July 6-7, 2002. Association for Computational Linguistics.

Levi, J. N. (1978). *The Syntax and Semantics of Complex Nominals*. New York: McGraw Hill.

Levow, Gina-Anne, Douglas W. Oard & Philip Resnik (2005). Dictionary-based techniques for cross-language information retrieval. *Information Processing and Management: an International Journal*, 41(3):523–547.

Liu, Hongfang, Alan R. Aronson & Carol Friedman (2002). A study of abbreviations in MEDLINE abstracts. In Isaac S. Kohane (Ed.), *AMIA 2002 – Proceedings of the Annual Symposium of the American Medical Informatics Association. Bio × medical Informatics: One Discipline*, pp. 464–468. San Antonio, TX, November 9-13, 2002. Philadelphia, PA: Hanley & Belfus.

Liu, Hongfang & Carol Friedman (2003). Mining terminological knowledge in large biomedical corpora. In Russ B. Altman, A. Keith Dunker, Lawrence Hunter, Tiffany A. Jung & Teri E. Klein (Eds.), *PSB 2003 – Proceedings of the Pacific Symposium on Biocomputing 2003*, pp. 415–426. Kauai, Hawaii, USA, January 3-7, 2003. Singapore: World Scientific Publishing.

Lovins, Julie B. (1968). Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11(1/2):22–31.

Lovis, Christian, Robert Baud, Pierre-André Michel & Jean-Raoul Scherrer (1997). Morphosemantems decomposition and semantic representation to allow fast and efficient natural language recognition. In Daniel R. Masys (Ed.), *AMIA'97 – Proceedings of the 1997 AMIA Annual Fall Symposium. The Emergence of*

'Internetable' Health Care: Systems that Really Work, p. 873. Nashville, TN, October 25-29, 1997. Philadelphia, PA: Hanley & Belfus. (extended version available on CD-ROM).

Lyman, Margaret, Naomi Sager, Led Tick, Ngo T. Nhan, Francois Borst & Jean-Raoul Scherrer (1991). The application of natural-language processing to healthcare quality assessment. *Medical Decision Making*, 11(4 Suppl.):65–68.

MacWhinney, Brian (1995). Language-specific prediction in foreign language learning. *Language Testing*, 12(3):292–320.

Manning, Christopher D. & Hinrich Schütze (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, MA; London, U.K.: Bradford Book & MIT Press.

Markó, Kornél, Robert Baud, Zweigenbaum Pierre, Lars Borin, Magnus Merkel & Stefan Schulz (2006a). Towards a multilingual medical lexicon. In *AMIA'06 – Proceedings of the 2006 Annual Symposium of the American Medical Informatics Association*, pp. 534–538. Washington, D.C., November 11-15, 2006. American Medical Informatics Association.

Markó, Kornél, Robert Baud, Pierre Zweigenbaum, Magnus Merkel, Maria Toporowska-Gronostaj, Dimitrios Kokkinakis & Stefan Schulz (2006b). Cross-lingual alignment of medical lexicons. In *LREC 2006 – Proceedings of the 5th International Conference on Language Resources and Evaluation Workshop: Acquiring and representing multilingual, specialized lexicons: the case of biomedicine*. Genua, Italy, May 24-26, 2006.

Markó, Kornél, Philipp Daumke, Jan Paetzold & Albrecht Zaiss (2006c). ICD-Coding mit MorphoSaurus. In *GMDS 2006 – Tagungsband der 51. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie*. Leipzig, Germany, September 10-14, 2006.

Markó, Kornél, Phillip Daumke, Stefan Schulz & Udo Hahn (2003). Cross-language MeSH indexing using morpho-semantic normalization. In Mark A. Musen

(Ed.), *AMIA'03 – Proceedings of the 2003 Annual Symposium of the American Medical Informatics Association. Biomedical and Health Informatics: From Foundations to Applications*, pp. 425–429. Washington, D.C., November 8-12, 2003. Philadelphia, PA: Hanley & Belfus.

Markó, Kornél, Udo Hahn, Stefan Schulz, Philipp Daumke & Percy Nohama (2004a). Interlingual indexing across different languages. In *RIAO 2004 – Conference Proceedings: Coupling Approaches, Coupling Media and Coupling Languages for Information Retrieval*, pp. 82–99. Avignon, France, 26-28 April 2004. Paris: Centre de Hautes Etudes Internationales d'Informatique Documentaire (CID).

Markó, Kornél, Stefan Schulz & Udo Hahn (2005a). Automatic lexicon acquisition for a medical cross-language information retrieval system. In *MIE 2005 – Proceedings of the XIX International Congress of the European Federation for Medical Informatics*, pp. 829–834. Geneva, Switzerland, August 28-31, 2005.

Markó, Kornél, Stefan Schulz & Udo Hahn (2005b). Automatische Generierung einer sprachübergreifenden Akronymdatenbank. In R. Klar, W Köpcke, K Kuhn, H. Lax, S. Weiland & A. Zaiss (Eds.), *GMDS 2005 – Tagungsband der 50. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie*, pp. 111–113. Freiburg, 11-15 September, 2005.

Markó, Kornél, Stefan Schulz & Udo Hahn (2005c). MorphoSaurus - design and evaluation of an interlingua-based, cross-language document retrieval engine for the medical domain. *Methods of Information in Medicine*, 44(4):537–545.

Markó, Kornél, Stefan Schulz & Udo Hahn (2005d). Multilingual lexical acquisition by bootstrapping cognate seed lexicons. In *RANLP 2005 – Proceedings of the International Conference on 'Recent Advances in Natural Language Processing'*, pp. 301–307. Borovets, Bulgaria, 21-23 September, 2005.

Markó, Kornél, Stefan Schulz & Udo Hahn (2005e). Unsupervised multilingual word sense disambiguation via an interlingua. In *AAAI 2005 – Proceedings of the 20th National Conference on Artificial Intelligence & IAAI'05 – Proceedings*

of the 17th Innovative Applications of Artificial Intelligence Conference, pp. 1075–1080. Pittsburgh, Pennsylvania, USA, July 9-13, 2004. Menlo Park, CA; Cambridge, MA: AAAI Press & MIT Press.

Markó, Kornél, Stefan Schulz & Udo Hahn (2006d). Automatic lexeme acquisition for a multilingual medical subword thesaurus. *International Journal of Medical Informatics*, 76(2-3):184–189.

Markó, Kornél, Stefan Schulz & Udo Hahn (2006e). Cross-lingual alignment of biomedical acronyms and their expansions. In *MIE 2006 – Proceedings of the 20th International Congress of the European Federation of Medical Informatics*, pp. 857–862. Maastricht, Netherlands, August 27 - 30, 2006. Amsterdam: IOS Press.

Markó, Kornél, Stefan Schulz, Alyona Medelyan & Udo Hahn (2005f). Bootstrapping dictionaries for cross-language information retrieval. In *SIGIR 2005 – Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 528–535. Salvador, Brazil, August 15-19, 2005. New York, NY: ACM.

Markó, Kornél, Stefan Schulz, Joachim Wermter, Michael Poprat & Udo Hahn (2004b). Cross-language document retrieval with MorphoSaurus. In E. Ammenwerth, W. Gaus, R. Haux, C. Lovis, K.P. Pfeiffer, B. Tilg & H.E. Wichmann (Eds.), *GMDS 2004 – Tagungsband der 49. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie*. Innsbruck, Austria.

McCarley, Jeffrey Scott (1999). Should we translate the documents or the queries in cross-language information retrieval? In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pp. 208–214. College Park, MD, USA, 20-26 June 1999. San Francisco, CA: Morgan Kaufmann.

McCray, Alexa T., Allen C. Browne & D. L. Moore (1988). The semantic structure of neo-classical compounds. In R. A. Greenes (Ed.), *SCAMC'88 – Proceedings*

of the 12th Annual Symposium on Computer Applications in Medical Care, pp. 165–168. Washington, D.C., November 1988. New York, N.Y.: EEE Computer Society Press.

McNamee, Paul & James Mayfield (2004). Character n-gram tokenization for european language text retrieval. *Information Retrieval*, 7(1-2):73–97.

Moulinier, Isabelle, J. Andrew McCulloh & Elizabeth Lund (2001). West group at CLEF 2000: Non-english monolingual retrieval. In *Cross-Language Information Retrieval and Evaluation: Workshop of Cross-Language Evaluation Forum, CLEF 2000, Lisbon, Portugal, September 21-22, 2000*, Vol. 2069, Lecture Notes in Computer Science. Heidelberg, Germany: Springer Verlag.

Namer, Fiammetta & Robert Baud (2005). Guessing lexical relations between biomedical terms: Towards a multilingual morphosemantics-based system. In *MIE 2005 – Proceedings of the 19th International Congress of the European Federation of Medical Informatics*. Geneva, Switzerland, August 28 - September 1, 2005. Amsterdam: IOS Press.

Nenadić, Goran, Irena Spasić & Sophia Ananiadou (2003). Terminology-driven mining of biomedical literature. *Bioinformatics*, 19(8):938–943.

Névéol, Aurélie, Vincent Mary, Arnaud Gaudinat, Célia Boyer, Alexandrina Rogozan & Stéfan J. Darmoni (2005a). A benchmark evaluation of the French MeSH indexers. In *AIME'05 — Proceedings of the 10th Conference on Artificial Intelligence in Medicine*, pp. 251–255.

Névéol, Aurélie, James G. Mork, Alan R. Aronson & Stéfan J. Darmoni (2005b). Evaluation of French and English MeSH indexing systems with a parallel corpus. In *AMIA'05 – Proceedings of the 2005 Annual Symposium of the American Medical Informatics Association*, pp. 565–569. Washington, D.C., November 22-26, 2005. Philadelphia, PA: Hanley & Belfus.

Ng, Hwee Tou & Hian Beng Lee (1996). Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *ACL'96 – Proceed-*

ings of the 34th Annual Meeting of the Association for Computational Linguistics, pp. 40–47. University of California at Santa Cruz, California, USA, 24-27 June 1996. San Francisco, CA: Morgan Kaufmann.

Norton, L. M. & Milos G. Pacak (1983). Morphosemantic analysis of compound word forms denoting surgical procedures. *Methods of Information in Medicine*, 22(1):29–36.

Nyström, M., M. Merkel, L. Ahrenberg, H. Petersson & H. Åhlfeld (2006). Creating a medical English-Swedish dictionary using interactive word alignment. *BMC Medical Informatics and Decision Making*, 6:35.

Oard, Douglas W. (2002). When you come to a fork in the road, take it: Multiple futures for CLIR research. In *SIGIR 2002 Workshop on the Future of Cross-Language Information Retrieval Research*. August 2002, Tampere, Finland.

Oard, Douglas W. & Anne R. Diekema (1998). Cross-language information retrieval. In Martha E. Williams (Ed.), *Annual Review of Information Science and Technology (ARIST), Vol. 33: 1998*, pp. 223–256. Medford, NJ: Information Today.

Okazaki, Naoaki & Sophia Ananiadou (2006). Clustering acronyms in biomedical text for disambiguation. In *LREC 2006 – Proceedings of the 5th International Conference on Language Resources and Evaluation*, pp. 959–962. Genua, Italy, May 24-26, 2006.

Pacak, Milos G., L. M. Norton & George S. Dunham (1980). Morphosemantic analysis of *-itis* forms in medical language. *Methods of Information in Medicine*, 19(2):99–105.

Park, Youngja & Roy J. Byrd (2001). Hybrid text mining for finding abbreviations and their definitions. In *EMNLP'01 – Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, pp. 126–133. Pittsburgh, PA.

Pirkola, Ari, Turid Hedlund, Heikki Keskustalo & Kalervo Järvelin (2001). Dictionary-based cross-language information retrieval: Problems, methods, and research findings. *Information Retrieval*, 4(3/4):209–230.

Pirkola, Ari, Heikki Keskustalo, Erkka Leppänen, Antti-Pekka Känsälä & Kalervo Järvelin (2002). Targeted s-gram matching: a novel n-gram matching technique for cross- and mono-lingual word form variants. *Information Research*, 7(2).

Popovič, Mirko & Peter Willett (1992). The effectiveness of stemming for natural-language access to Slovene textual data. *Journal of the American Society for Information Science*, 43(5):384–390.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137.

Pouliquen, Bruno, Ralf Steinberger & Camelia Ignat (2003). Automatic identification of document translations in large multilingual document collections. In *RANLP 2003 – Proceedings of the International Conference on 'Recent Advances in Natural Language Processing'*, pp. 401–408. Borovets, Bulgaria, 8-9 September 2003.

Pratt, Arnold W. & Milos G. Pacak (1969). Identification and transformation of terminal morphemes in medical English. *Methods of Information in Medicine*, 8(2):84–90.

Pustejovsky, James, José Castaño, Brent Cochran, Maciej Kotecki & Michael Morrell (2001). Automatic extraction of acronym-meaning pairs from MEDLINE databases. In V. L. Patel, R. Rogers & R. Haux (Eds.), *MEDINFO 2001 – Proceedings of the 10th World Congress on Medical Informatics. Vol. 1*, Studies in Health Technology and Informatics 84, pp. 371–375. London, U. K., September 2001. Amsterdam: IOS Press.

Rapp, Reinhard (1999). Automatic identification of word translations from unrelated English and German corpora. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pp. 519–526. College Park, MD, USA, 20-26 June 1999. San Francisco, CA: Morgan Kaufmann.

Rector, Alan L. (1999). Clinical terminology: Why is it so hard? *Methods of Information in Medicine*, 38:239–252.

Rector, Alan L., Sean Bechhofer, Carole A. Goble, Ian Horrocks, W. Anthony Nowlan & W. Danny Solomon (1997). The GRAIL concept modelling language for medical terminology. *Artificial Intelligence in Medicine*, 9(2):139–171.

Ribeiro, António, Gaël Dias, Gabriel Lopes & João Mexia (2001). Cognates alignment. In *Proceedings of Machine Translation Summit VIII*, pp. 287–293. Santiago de Compostela, Spain, September 18-22, 2001.

Roche (2003). *Roche Lexikon Medizin* (5th ed.). Urban and Fischer.

Romacker, Martin & Udo Hahn (2001). Coping with different types of ambiguity using a uniform context handling mechanism. In *Natural Language Processing and Information Systems. Revised Papers of the 5th International Conference on Applications of Natural Language to Information System – NLDB 2000*, Vol. 1959, Lecture Notes in Computer Science, pp. 42–53. Versailles, France, June 28-30, 2000. Berlin: Springer.

Romacker, Martin, Katja Markert & Udo Hahn (1999). Lean semantic interpretation. In *IJCAI'99 – Proceedings of the 16th International Joint Conference on Artificial Intelligence*, Vol. 2, pp. 868–875. Stockholm, Sweden, July 31 - August 6, 1999. San Francisco, CA: Morgan Kaufmann.

Rose, Tony, Mark Stevenson & Miles Whitehead (2002). The REUTERS Corpus Volume 1: From yesterday's news to tomorrow's language resources. In M.G. Rodriguez & C. Paz Suarez Araujo (Eds.), *LREC 2002 – Proceedings of the 3rd International Conference on Language Resources and Evaluation. Vol. 3*, pp. 827–833. Las Palmas de Gran Canaria, Spain, 29-31 May, 2002. Paris: European Language Resources Association (ELRA).

Rosemblat, Graciela, Darren Gemoets, Allen C. Browne & Tony Tse (2003). Machine translation-supported cross-language information retrieval for a consumer health resource. In Mark A. Musen (Ed.), *AMIA'03 – Proceedings of the 2003*

Annual Symposium of the American Medical Informatics Association. Biomedical and Health Informatics: From Foundations to Applications, pp. 564–568. Washington, D.C., November 8-12, 2003. Philadelphia, PA: Hanley & Belfus.

Rosemblat, Graciela & Laurel Graham (2006). Cross-language search in a monolingual health information system: Flexible designs and lexical processes. In *ISKO'06 – Proceedings of the 9th International Society for Knowledge Organization Conference*, pp. 173–182. Vienna, Austria, July 2006.

Rowe, Raymond C. (2003). Abbreviation mania and acronymical madness. *Drug Discovery Today*, 8(16):732–733.

Russell, Graham J., Graeme D. Ritchie, Stephen G. Pulman & Alan W. Black (1986). A dictionary and morphological analyzer for English. In *COLING '86 – Proceedings of the 11th International Conference on Computational Linguistics*, pp. 277–279. Bonn, Germany, August 25-29, 1986. Bonn: Institut für angewandte Kommunikations- und Sprachforschung (IKS).

Sager, Naomi, Carol Friedman & Margaret S. Lyman (Eds.) (1987). *Medical Language Processing. Computer Management of Narrative Text*. Reading, MA: Addison-Wesley.

Sager, Naomi, Margaret Lyman, Christine E. Bucknall, Ngo Thanh Nhan & Leo J. Tick (1994). Natural language processing and the representation of clinical data. *Journal of the American Medical Informatics Association*, 1(2):142–160.

Salton, Gerald (Ed.) (1971). *The* SMART *Retrieval System: Experiments in Automatic Document Processing*. Englewood Cliffs, NJ: Prentice-Hall.

Salton, Gerald & Chris Buckley (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523.

Savoy, Jacques (2003a). Cross-language information retrieval: experiments based on CLEF 2000 corpora. *Information Processing and Management: an International Journal*, 39(1):75–115.

Savoy, Jacques (2003b). Report of CLEF-2003 multilingual tracks. In Carol Peters (Ed.), *Working Notes for the 2003 CLEF Workshop*. Trondheim, Norway, 21-22 August.

MeSH (2005). *Medical Subject Headings*. Bethesda, MD: National Library of Medicine.

OPS (2006). *Operationen- und Prozedurenschlüssel*. Deutscher Ärzte Verlag.

Schiller, Anne & Petra Steffens (1991). Morphological processing in the two-level paradigm. In O. Herzog & C.-R. Rollinger (Eds.), *Text Understanding in* LILOG*. Integrating Computational Linguistics and Artificial Intelligence. Final Report on the* IBM *Germany* LILOG*-Project*, Lecture Notes in Artificial Intelligence 546, pp. 112–126. Berlin: Springer.

Schulz, Stefan & Udo Hahn (2000). Morpheme-based, cross-lingual indexing for medical document retrieval. *International Journal of Medical Informatics*, 59(3).

Schulz, Stefan, Martin Honeck & Udo Hahn (2002). Biomedical text retrieval in languages with a complex morphology. In Stephen Johnson (Ed.), *Proceedings of the ACL/NAACL 2002 Workshop on 'Natural Language Processing in the Biomedical Domain'*, pp. 61–68. University of Pennsylvania, Philadelphia, PA, USA, July 11, 2002. New Brunswick, NJ: Association for Computational Linguistics (ACL).

Schulz, Stefan, Kornél Markó, Philipp Daumke, Udo Hahn, Susanne Hanser, Percy Nohama, Roosewelt Leite de Andrade, Edson Pacheco & Martin Romacker (2006). Semantic atomicity and multilinguality in the medical domain: Design considerations for the MorphoSaurus subword lexicon. In *LREC 2006 – Proceedings of the 5th International Conference on Language Resources and Evaluation*. Genua, Italy, May 24-26, 2006.

Schulz, Stefan, Kornél Markó, Eduardo Sbrissia, Percy Nohama & Udo Hahn (2004). Cognate mapping: A heuristic strategy for the semi-supervised acquisition of a Spanish lexicon from a Portuguese seed lexicon. In *COLING Geneva 2004 –*

Proceedings of the 20th International Conference on Computational Linguistics, Vol. 2, pp. 813–819. Geneva, Switzerland, August 23-27, 2004. Association for Computational Linguistics.

Schütze, Hinrich (1992). Dimensions of meaning. In *Supercomputing'92 – Proceedings of the 1992 ACM/IEEE Conference on Supercomputing*, pp. 787–796. Minneapolis, MN, 1992. IEEE Computer Society Press.

Schwartz, Ariel S. & Marti A. Hearst (2003). A simple algorithm for identifying abbreviation definitions in biomedical text. In Russ B. Altman, A. Keith Dunker, Lawrence Hunter, Tiffany A. Jung & Teri E. Klein (Eds.), *PSB 2003 – Proceedings of the Pacific Symposium on Biocomputing 2003*, pp. 451–462. Kauai, Hawaii, USA, January 3-7, 2003. Singapore: World Scientific Publishing.

Sebastiani, Fabrizio (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47.

Shatkay, Hagit & Ronen Feldman (2003). Mining the biomedical literature in the genomic era: An overview. *Journal of Computational Biology*, 10(6):821–855.

Simpson, John & Edmund Weiner (1989). *The Oxford English Dictionary* (2nd ed.). Oxford University Press.

Sproat, Richard (1992). *Morphology and Computation*. Cambridge, MA: MIT Press.

Spyns, Peter (1996). Natural language processing in medicine: An overview. *Methods of Information in Medicine*, 35(4/5):285–301.

Taber (2005). *Taber's Cyclopedic Medical Dictionary* (20th Rev ed.). F. A. Davis Company.

Tellex, Stefanie, Boris Katz, Jimmy J. Lin, Aaron Fernandes & Gregory Marton (2003). Quantitative evaluation of passage retrieval algorithms for question answering. In *SIGIR 2003 – Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 41–47. Toronto, Canada, July 28 - August 1, 2003. ACM.

Toman, J. (1987). *Wortsyntax. Eine Diskussion ausgewählter Probleme deutscher Wortbildung*. Tübingen: Max Niemeyer.

Tomlinson, Stephen (2001). Stemming evaluated in 6 languages by hummingbird searchserver$^{\text{tm}}$ at CLEF 2001. In Carol Peters, Martin Braschler, Julio Gonzalo & Michael Kluck (Eds.), *Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Darmstadt, Germany, September 3-4, 2001*, Vol. 2406, pp. 278–287.

Tordai, Anna & Maarten de Rijke (2005). Four stemmers and a funeral: Stemming in Hungarian at CLEF 2005. In Carol Peters (Ed.), *Working Notes for the 2005 CLEF Workshop*. Vienna, Austria, 21-23 September.

Towell, Geoffrey & Ellen M. Voorhees (1998). Disambiguating highly ambiguous words. *Computational Linguistics*, 24(1):125–145.

Trost, Harald (1990). The application of two-level morphology to non-concatenative German morphology. In *COLING'90 – Papers Presented at the 13th International Conference on Computational Linguistics on the Occasion of the 25th Anniversary of COLING & the 350th Anniversary of Helsinki University*, Vol. 2, pp. 371–376. Helsinki, Finland, 1990.

Trost, Harald (1993). Coping with derivation in a morphological component. In *EACL'93 – Proceedings of the 6th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 368–376. Utrecht, The Netherlands, 21-23 April 1993. Association for Computational Linguistics.

Turcato, Davide (1998). Automaticaly creating bilingual lexicons for machine translation from bilingual text. In *COLING/ACL'98 – Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics & 17th International Conference on Computational Linguistics*, Vol. 2, pp. 1299–1306. Montréal, Quebec, Canada, August 10-14, 1998.

UMLS (2005). *Unified Medical Language System*. Bethesda, MD: National Library of Medicine.

UNESCO (1995). *A Structured List of Descriptors for Indexing and Retrieving Literature in the Fields of Education, Science, Social and Human Science, Culture, Communication and Information.* Paris: UNESCO Publishing.

Volk, Martin, Bärbel Ripplinger, Spela Vintar, Paul Buitelaar, Diana Raileanu & Bogdan Sacaleanu (2002). Semantic annotation for concept-based cross-language medical information retrieval. *International Journal of Medical Informatics*, 67(1/3):79–112.

Voorhees, Ellen M. (1993). Using WordNet to disambiguate word senses for text retrieval. In R. Korfhage, E. Rasmussen & P. Willett (Eds.), *SIGIR'93 – Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 171–180. Pittsburgh, PA, USA, June 27 - July 1, 1993. New York, NY: ACM.

Vossen, Piek (Ed.) (1998). EuroWordNet*: A Multilingual Database with Lexical Semantic Networks.* Dordrecht: Kluwer Academic Publishers.

Waegemann, C.P. (1996). The five levels of electronic health records. *M.D. Computing: Computers in Medical Practice*, 13(3):199–203.

Waegemann, C.P. (2002). The vision of electronic health records. *The Journal of Medical Practice Management*, 8(2):63–65.

Weeber, Marc, James G. Mork & Alan R. Aronson (2001). Developing a test collection for biomedical word sense disambiguation. In Suzanne Bakken (Ed.), *AMIA 2001 – Proceedings of the Annual Symposium of the American Medical Informatics Association*, pp. 746–750. Washington, D.C., November 3-7, 2001. Philadelphia, PA: Hanley & Belfus.

Wermter, Joachim & Udo Hahn (2004). Really, is medical sublanguage that different? Experimental counter-evidence from tagging medical and newspaper corpora. In Marius Fieschi, Enrico Coiera & Yu-Chan Jack Li (Eds.), *MEDINFO 2004 – Proceedings of the 11th World Congress on Medical Informatics. Vol. 1*, pp. 560–564. San Francisco, CA, USA, September 7-11, 2004.

Weske-Heck, Gesa, Albrecht Zaiss, Stefan Schulz, Wolfgang Giere, Michael Schopen & Rüdiger Klar (2002). The German Specialist Lexicon. In Isaac S. Kohane (Ed.), *AMIA 2002 – Proceedings of the Annual Symposium of the American Medical Informatics Association. Biomedical Informatics: One Discipline*, pp. 884–888. San Antonio, TX, November 9-13, 2002.

Widdows, Dominic, Beate Dorow & Chiu-Ki Chan (2002). Using parallel corpora to enrich multiligual lexical resources. In M.G. Rodriguez & C. Paz Suarez Araujo (Eds.), *LREC 2002 – Proceedings of the 3rd International Conference on Language Resources and Evaluation. Vol. 1*, pp. 240–245. Las Palmas de Gran Canaria, Spain, 29-31 May, 2002.

Wingert, F. (1977). Morphosyntaktische Zerlegung von Komposita der medizinischen Sprache. *Methods of Information in Medicine*, 16(4):248–255.

Wingert, F. (1985). Morphologic analysis of compound words. *Methods of Information in Medicine*, 24(3):155–162.

Wolff, Susanne (1984). The use of morphosemantic regularities in the medical vocabulary for automatic lexical coding. *Methods of Information in Medicine*, 23(4):195–203.

Wren, Jonathan D., Jeffrey T. Chang, James Pustejovsky, Eytan Adar, Harold R. Garner & Russ B. Altman (2005). Biomedical term mapping databases. *Nucleic Acids Research*, 33(1):D289–293.

Wren, Jonathan D. & Harold R. Garner (2002). Heuristics for identification of acronym-definition patterns within text: Towards an automated construction of comprehensive acronym-definition dictionaries. *Methods of Information in Medicine*, 41(5):426–434.

Yarowsky, David (1992). Word-sense disambiguation using statistical models of ROGET's categories trained on large corpora. In *COLING'92 – Proceedings of the 14th International Conference on Computational Linguistics*, pp. 454–460. Nantes, France, 23-28 August 1992. ICCL.

Yarowsky, David & Richard Wicentowski (2000). Minimally supervised morphological analysis by multimodal alignment. In *ACL'00 – Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pp. 207–216. Hong Kong, 1-8 August 2000. San Francisco, CA: Morgan Kaufmann.

Zeng, Q. & J.J. Cimino (1996). Mapping medical vocabularies to the Unified Medical Language System. In *AMIA'96 – Proc. of the 1996 AMIA Annual Fall Symposium*, pp. 105–109.

Zhang, Ying & Phil Vines (2004). Using the web for automated translation extraction in cross-language information retrieval. In *SIGIR 2004 – Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 162–169. Sheffield, United Kingdom.

Zweigenbaum, Pierre, Robert Baud, Anita Burgun, Fiammetta Namer, Éric Jarrousse, Natalia Grabar, Patrick Ruch, Franck Le Duff, Jean-François Forget, Magaly Douyère & Stéfan Darmoni (2005). UMLF – a unified medical lexicon for French. *International Journal of Medical Informatics*, 74(2/4):119–124.

Zweigenbaum, Pierre, Jacques Bouaud, Bruno Bachimont, Jean Charlet & Jean-Francois Boisvieux (1997). Evaluating a normalized conceptual representation produced from natural language patient discharge summaries. In R. Masys (Ed.), *AMIA'97 – Proc. of the 1997 AMIA Annual Fall Symposium*, pp. 590–594. Nashville, TN, October 25-29, 1997. Philadelphia, PA: Hanley & Belfus.

Zweigenbaum, Pierre, Stéfan J. Darmoni & Natalia Grabar (2001). The contribution of morphological knowledge to French MeSH mapping for information retrieval. In Suzanne Bakken (Ed.), *AMIA 2001 – Proceedings of the Annual Symposium of the American Medical Informatics Association*, pp. 796–800. Washington, D.C., November 3-7, 2001. Philadelphia, PA: Hanley & Belfus.